

Masterarbeit zur Erlangung des akademischen Grades
Master of Arts UZH
der Philosophischen Fakultät der Universität Zürich
im Rahmen des fakultätsübergreifenden Master-Studiengangs
Computerlinguistik und Sprachtechnologie

Massenüberwachung mittels Computerlinguistik und Sprachtechnologie im Lichte der Snowden-Enthüllungen

Recherchen zum Stand der Technik und kritisch-illustratives Datenprojekt zur automatischen Generierung von Selektoren für verdachtsunabhängige Massenüberwachung von Datenströmen mit manueller Evaluation resultierender Treffer

Verfasser: Hernâni Marques Madeira

{h2m@access.uzh,hernani@ccc-ch}.ch
7FE5 71F9 3B0C AE18 8424 C7C2 7B83 6E41 F7AB 9CE5
hernani@jabber.ccczh.ch (OTR-verschlüsselt)

Referent: Prof. Dr. Martin Volk

Betreuer: Dr. Noah Bubenhofer

Institut für Computerlinguistik

Abgabedatum: 30.11.2015

PUBLIC//V1.14//20160902

V1.14: EDITION NDG-ABSTIMMUNG[1]



[1]Formal sind diese Bilder und der Aufruf natürlich nicht Teil der offiziellen Abgabe.



An Schweizer Stimmberechtigte: Am 25. September NEIN zum NDG!

https://www.stopndb.ch/

Abstract

The master thesis gives an overview of the state-of-the-art in mass surveillance methods and practices following the Snowden relevations and scientific papers with a focus on NLP methods. For the practical part of this work, genesis and the nature of search terms (content or soft selectors) used in finding text material by the needle-in-a-haystock methodology are researched, implemented, evaluated and critically dicscussed. Quantitatively and technically, as expected, it is found that false positive rates in general are high, eventhough some of the surveillance methods assumed perform better considering specific cases in comparison. Qualitatively and ethically, it is found such methods in action can never be – socially – proportionate considering (1) their lack to take into account the semantic complexity of natural language and (2) by the incentives they create to gather more and more data in order to reduce possible sparse data problems or to produce true positives (at all) – thus increasingly violating basic privacy rights.

Zusammenfassung

Die Masterarbeit bietet eine Übersicht vom Stand der Technik der Massenüberwachung, gegeben die Snowden-Enthüllungen und wissenschaftliche Papers mit einem Fokus auf CL-Methoden. Im praktischen Teil der Arbeit wird die Entstehung und Natur inhaltlicher Suchbegriffe oder Selektoren, welche zur Suche von Textmaterial entsprechend der Nadel-im-Heuhaufen-Methodologie eingesetzt werden, erforscht, implementiert, evaluiert und kritisch diskutiert. Wie erwartet, wird quantitativ und technisch befunden, dass die Rate der False-Positives generell hoch ausfällt, wenn auch einige der angenommenen Überwachungsmethoden in spezifischen Fällen im Vergleich besser funktionieren. Qualitativ und ethisch wird befunden, dass der Einsatz solcher Methoden niemals – sozial – verhältnismässig sein kann, weil sie (1) der semantischen Komplexität von natürlicher Sprache nicht gerecht werden und (2) Anreize dafür geschaffen werden, immer mehr Daten zu sammeln, um mögliche Sparse-Data-Probleme zu reduzieren oder (überhaupt) True-Positives zu produzieren – zum Preis grundlegende Rechte der Privatsphäre immer mehr zu untergraben.

Danksagung

Meinen grössten Dank gebührt meinen beiden Schätzen Alessandra und Sabrina für die motivationale Unterstützung, die ich während der ganzen Zeit an meiner Masterarbeit geniessen konnte: sie war vital, zu einem Ende zu kommen. Sabrina sei insbesondere dafür gedankt, nicht nur in meinem Leben, sondern auch im Rahmen der Evaluation der Ergebnisse dieser Arbeit als meine bessere Hälfte operiert zu haben.
Ich danke auch meinen Eltern, mich ganz bis zum Schluss meines Bildungsweges, mithin hiermit zur Hochschulebene – so auch während der Masterarbeit – in finanziellen Belangen zu stützen. Ich weiss, dass das ein Privileg ist.

Weiterhin bin ich besonders meinem Betreuer Dr. Noah Bubenhofer dankbar dafür, eingangs nicht nur völlig unproblematisch zugesagt zu haben, für die Thematik dieser Arbeit mein Betreuer zu sein, sondern auch dafür, mich ohne jeden Druck und doch mit Bestimmtheit durch diese Masterarbeit hindurch orientiert zu haben. Genauso sei Prof. Dr. Martin Volk (als Referenten) gedankt, das Thema bewilligt und den letzten, finalen Blick über die Arbeit zu werfen.

Auch sei dem ganzen Institut für Computerlinguistik und im Allgemeinen der Universität Zürich gedankt, für die Lehre, die ich über die Jahre erfahren habe und die Forschungseinsichten, die ich gewinnen konnte und die in der einen oder anderen Form – wenn auch bloss subtil – in dieser Arbeit ihren Niederschlag finden.

Abschliessend möchte ich allen Personen meinen Dank aussprechen, die sich tagtäglich dafür einsetzen, dass Grundrechte auch im Internetzeitalter mit allem Elan verteidigt werden und dafür kämpfen, dass ein mündiger und rascher Austritt aus den mittelalterlichen Zuständen der Verfassungs- und Rechtelosigkeit im Cyberspace erfolgt. Ich danke ganz besonders allen Menschen, die ohne auf Lorbeeren aus zu sein, politisch und technisch aktivistisch tätig sind, wie solche (der Netzwerke) von und um Anonymous; auch danke ich allen, die dies namentlich in einer Weise tun, dass ihr Leben wesentlich (negativ) davon betroffen wird – wie Edward Snowden; zu guter Letzt bedenke ich auch verstorbenen Pionieren wie den durch massiven Druck der US-Justiz in den Tod getriebenen Aktivisten und Hacker Aaron Swartz.

Inhaltsverzeichnis

Abstract **i**

Danksagung **ii**

Inhaltsverzeichnis **iii**

Tabellenverzeichnis **vii**

Abkürzungsverzeichnis **viii**

1 Einführung **1**

1.1 Motivation und Sinn . . . 1
1.2 Forschungsfragen und Hypothesen . . . 4
1.3 Fokus und Grenzen . . . 5
1.4 Aufbau . . . 7

2 Stand der Technik des Überwachungskomplexes **9**

2.1 Zu Wesen und Spielarten der Massenüberwachung . . . 10
2.1.1 Was ist Überwachung und spezifisch Massenüberwachung? . . . 10
2.1.2 Welche Ansatzpunkte zur digitalen Überwachung gibt es? . . . 13
2.1.2.1 Überwachung von Datenströmen . . . 13
2.1.2.2 Ausleitung Daten Endgerät (an beiden Enden) . . . 14
2.1.2.3 Ausleitung Daten Servergerät (bei Client-Server-Modellen) . . . 14
2.1.3 Beispiele: Fälle und Folgen von Verdächtigungen auf linguistischer Basis 15
2.1.3.1 Schweiz: “Rütli-Bomber” ungrammatikalisch in seiner Rache . . . 15
2.1.3.2 Deutschland: Andrej Holm als Soziologe zum “Terroristen” . . . 16
2.1.3.3 USA: “Unabomber” Theodore Kaczynski und seine Phrase . . . 17
2.2 Computerlinguistische Forschung zur Ausübung von Massenüberwachung . 17
2.2.1 Rolle der Computerlinguistik für die Massenüberwachung . . . 17
2.2.2 EU: INDECT-Projekt . . . 19
2.2.3 USA: HLT-Programm der NSA . . . 20
2.2.4 USA: NSA-Patente mit Computerlinguistik-Bezug . . . 21
2.3 Im Fokus: Sprachtechnologische Produkte zur Massenüberwachung . . . 22
2.3.1 Snowden-Enthüllungen: XKeyscore . . . 22
2.3.2 WikiLeaks-Publikation: Search & Detect . . . 24
2.4 Befunde: Logiken der Massenüberwachung auf Basis von Selektoren . . . 26
2.4.1 Was sind Selektoren? . . . 26

2.4.2 FVEY: Echelon . . . 27
2.4.3 Deutschland: NSAUA zur Massenüberwachung des BND . . . 30
2.4.3.1 Zahl von Selektoren . . . 30
2.4.3.2 Graulich-Bericht . . . 30
2.4.3.3 Spionage von Freunden und “hochkritische” Selektoren . . . 32
2.4.4 Schweiz: Onyx oder Massenüberwachung von Militär und NDB . . . 32
2.4.4.1 Untersuchungen der GPDel zur Funkaufklärung . . . 34
2.4.4.2 Geplante Kabelaufklärung im NDG . . . 35
2.5 Befunde: Konkrete Selektoren zur Massenüberwachung . . . 37
2.5.1 FVEY: Mögliche Echelon-Selektorenlisten . . . 37
2.5.2 FVEY: XKeyscore und Beispiele von Selektoren . . . 38
2.5.3 USA: WikiLeaks-Publikationen . . . 39
2.5.4 Deutschland: Medienberichte und NSAUA . . . 39
2.5.5 Schweiz: Medienberichte und Onyx . . . 39

3 Automatische Generierung von Selektoren 41
3.1 Grundannahmen . . . 41
3.2 Trainingsdaten . . . 42
3.2.1 Auswahlkriterien . . . 42
3.2.2 Aufbereitung . . . 43
3.2.3 Die Trainingskorpora Aufbau und PNOS . . . 43
3.3 Methode 1: TFIDF-Modell . . . 44
3.3.1 Grundlagen . . . 44
3.3.2 Einzelworte . . . 45
3.3.3 Wort-2-Gramme . . . 45
3.3.4 Wortkombinationen von 2–5 Worten . . . 45
3.4 Methode 2: Verdachtssprache-Modell . . . 46
3.4.1 Grundlagen . . . 47
3.4.2 Intensivierende Wortkombinationen . . . 48
3.4.3 Skandalisierende Wortkombinationen . . . 49
3.4.4 Wortkombinationen mit Verschwörungsvokabular . . . 49
3.5 Methode 3: LDA-Modell . . . 49
3.5.1 Grundlagen . . . 49
3.5.2 Kombinationen zu 2, 3, 5, 8 und 10 Worten . . . 50
3.6 Selektorenauswahl für die Methoden 1 und 2 . . . 51

4 Evaluation der Selektoren 52
4.1 Zum Umfang der Evaluation . . . 52
4.2 Die Evaluationssysteme und -daten im Einzelnen . . . 53
4.2.1 Whitenet-Index: DuckDuckGo . . . 53
4.2.2 Whitenet-Index: StartPage . . . 53
4.2.3 Darknet-Index: Not Evil . . . 54
4.2.4 P2P-Index: YaCy . . . 54

4.2.5 Privat-Index: Eigener 10-Tages-Datenstrom “Own” 55
4.3 Manuelle Annotation 56
4.3.1 Durchführung der manuellen Annotation 56
4.3.2 Statistik der manuellen Annotation 59
4.4 Ergebnisse 60
4.4.1 Trefferstatistik 60
4.4.1.1 Nach Evaluationssystem 61
4.4.1.2 Nach Überwachungsmodell 61
4.4.2 True-Positive-Statistik 62
4.4.2.1 Nach Evaluationssystem 63
4.4.2.2 Nach Überwachungsmethode 63
4.4.2.3 Precision im Durchschnitt: Top-10 64
4.4.2.4 Score im Durchschnitt: Top-10 64
4.4.2.5 Precision+Score im Durchschnitt: Top-10 65

5 Schlussbetrachtungen **66**
5.1 Diskussion des Versuchs und der Ergebnisse 66
5.1.1 Zur Güte der Evaluationsergebnisse 66
5.1.2 Willkürpotenzial in der Auswahl des Trainingsmaterials 67
5.1.3 Willkürpotenzial in der Generierung der Selektoren 67
5.1.4 Offener Spielraum bei der Interpretation der Treffer 68
5.1.5 Verhältnismässigkeit und Massenüberwachung 68
5.2 Strategien im Umgang mit der Massenüberwachung 69
5.2.1 Chilling Effects: Freiheit durch Anpassung? 70
5.2.2 Anonymisierung und Obfuskation der Urheberschaft 71
5.2.3 Anonymisierung und Verschlüsselung der Daten(-Wege) 71
5.3 Fazit 72

Glossar **75**

Quellenverzeichnis **76**

Selbstständigkeitserklärung **89**

A Tabellen **90**
A.1 Selektoren nach Modell 90
A.1.1 TFIDF 90
A.1.2 Verdachtssprache 91
A.1.3 LDA 92

B Abstracts computerlinguistisch relevanter NSA-Patente **94**
B.1 Method of retrieving documents that concern the same topic (1995) 94
B.2 Language-independent method of generating index terms (1998) 94

B.3 Automatically generating a topic description for text and searching and sorting text by topic using the same (1999) . . . 95
B.4 Device and method for full-text large-dictionary string matching using n-gram hashing (2001) . . . 95
B.5 Method for finding large numbers of keywords in continuous text streams (2001) 95
B.6 Method of summarizing text using just the text (2005) . . . 96
B.7 Method of summarizing text by sentence extraction (2006) . . . 96
B.8 Method of optical character recognition using feature recognition and baseline estimation (2008) . . . 96
B.9 Natural language database searching using morphological query term expansion (2010) . . . 97
B.10 Method of database searching (2010) . . . 97
B.11 Method of identifying topic of text using nouns (2010) . . . 98
B.12 Method of assessing language translation and interpretation (2012) . . . 98
B.13 Device for and method of language processing (2013) . . . 98

C Inoffizielles oder unpubliziertes Quellenmaterial 100
C.1 XKeyscore-Regeln im Zusammenhang mit Tor-Anonymisierungstechnologien 100
C.2 Daniel Mossbrucker (2015): Digitale Informantenschutzrechte . . . 103

D Code 117
D.1 Aufbereitung Daten und Erzeugung Selektoren . . . 117
D.1.1 “Own”-Datenformate nach file(1)-Analyse: Top 50 (Originaldateien) . . . 117
D.1.2 “Own”-Datenformate nach file(1)-Analyse: Top 50 (Keine Duplikate) . . . 118

E Daten 120
E.1 Trainingsdaten . . . 120
E.2 Evalautionsdaten . . . 120

F Hashsummen 121
F.1 Code . . . 121
F.2 Daten . . . 121

Tabellenverzeichnis

2.1 Übersicht INDECT-Forschung: Methoden der Computerlinguistik 19
3.1 Deskriptive Statistik: Korpora *Aufbau* und *PNOS* 44
3.2 Verdachtssprache-Modell: Intensivierer, Skandalisierer und Verschwörungsvokabular 48
4.1 Manuelle Annotation H und S 59
4.2 Erwartungswerte H und S für "Links-" (L) und "Rechtsextremismus" (R) . . . 60
4.3 Kappa-Werte für "Links-" (L) und "Rechtsextremismus" (R) 60
4.4 Trefferstatistik: Nach Evaluationssystem 61
4.5 Trefferstatistik: Nach Überwachungsmodell 61
4.6 True-Positive-Statistik: Precision- und Score-Werte nach Evaluationsdaten . 63
4.7 True-Positive-Statistik: Precision- und Score-Werte nach Überwachungsmodell 63
4.8 Precision im Durchschnitt: Top-10 64
4.9 Score im Durchschnitt: Top-10 65
4.10 Precision+Score im Durchschnitt: Top-10 65
A.1 Selektoren TFIDF-Modell: Einzelworte 90
A.2 Selektoren TFIDF-Modell: Wort-2-Gramme 90
A.3 Selektoren TFIDF-Modell: 2–5-Wortkombationen 90
A.4 Selektoren Verdachtssprache-Modell: intensivierend (generell) 91
A.5 Selektoren Verdachtssprache-Modell: intensivierend (absolut) 91
A.6 Selektoren Verdachtssprache-Modell: intensivierend (hoch) 91
A.7 Selektoren Verdachtssprache-Modell: intensivierend (extrem hoch) 91
A.8 Selektoren Verdachtssprache-Modell: Verschwörungsvokabular 92
A.9 Selektoren Verdachtssprache-Modell: skandalisierend 92
A.10 Selektoren LDA-Modell: 2-Wort-Topics 92
A.11 Selektoren LDA-Modell: 3-Wort-Topics 92
A.12 Selektoren LDA-Modell: 5-Wort-Topics 93
A.13 Selektoren LDA-Modell: 8-Wort-Topics 93
A.14 Selektoren LDA-Modell: 10-Wort-Topics 93

Abkürzungsverzeichnis

BfV	Bundesamt für Verfassungsschutz [Deutschland]
BND	Bundesnachrichtendienst [Deutschland]
BÜPF	Bundesgesetz zur Überwachung des Post- und Fernmeldeverkehrs [Schweiz]
CA	Certification Authority
CCC	Chaos Computer Club
COMINT	Communications Intelligence
DAP	Dienst für Analyse und Prävention [Schweiz]
EDÖB	Eidgenössischer Datenschutz- und Öffentlichkeitsbeauftragter
EMRK	Europäische Menschenrechtskonvention [Konvention zum Schutz der Menschenrechte und Grundfreiheiten]
FVEY	Five Eyes
FP7	Framework Programme 7 [EU]
GCHQ	Government Communications Headquarters [UK]
GPDel	Geschäftsprüfungsdelegation [Schweiz]
HLT	Human Language Technology
HUMINT	Human Intelligence
IMINT	Image Intelligence
INDECT	Intelligent information system supporting observation, searching and detection for security of citizens in urban environment [EU]
KI	Künstliche Intelligenz
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
LSI	Latent Semantic Indexing
NDB	Nachrichtendienst des Bundes [Schweiz]
NDG	Nachrichtendienstgesetz [Bundesgesetz über den Nachrichtendienst [Schweiz]]
NER	Named Entity Recognition
NIPF	National Intelligence Priorities Framework [NIPF]
NSA	National Security Agency [USA]
NSAUA	NSA-Untersuchungsausschuss [Erster parlamentarischer Untersuchungsausschuss des 18. Bundestages [Deutschland]]

OCR	Optical Character Recognition
OSINT	Open Source Intelligence
PNOS	Partei National Orientierter Schweizer
SIGINT	Signals Intelligence
SND	Strategischer Nachrichtendienst [Schweiz]
STOA	Science and Technology Options Assessment [EU]
STT	Speech-To-Text
TA	Topic Analysis
TAO	Tailored Access Operations [NSA]
TLS	Transport Layer Security
TPP	[NSA] Technology Transfer Program
TFIDF	Term Frequency – Inverse Document Frequency
UKI	Unabhängige Kontrollinstanz [Schweiz]
UKUSA	United Kingdom – United States of America Agreement
VDS	Vorratsdatenspeicherung
VEKF	Verordnung über die elektronische Kriegführung und die Funkaufklärung [Schweiz]
XKS	XKeyscore
ZNDG	Bundesgesetz über die Zuständigkeiten im Bereich des zivilen Nachrichtendienstes [Schweiz]

1 Einführung

Wie ein langsames Gift sickert in die Gesellschaft, dass man sich sowohl mit Massenüberwachung als auch mit andauernder Spionage gegen die Bevölkerung und die Spitzen von Politik und Wirtschaft einrichtet. Die Ohnmacht, die aus dem Unwillen der politischen Eliten rührt, dem in- und ausländischen Geheimdienstsumpf etwas entgegenzusetzen, wird zum Normalzustand. Geheimdienste spitzeln eben, wie sie wollen, lautet das hilflose Glaubensbekenntnis.

– Constanze Kurz[2] am 13. Juli 2015 [57]

1.1 Motivation und Sinn

Spätestens[3] seit dem Anbeginn der Snowden-Enthüllungen im Juni 2013 wird deutlich, wie weit fortgeschritten der weltweite Überwachungskomplex aus Privatindustrie und Geheimdiensten ist. Speziell die Enthüllungen um die Systeme *PRISM*, *Tempora* und *XKeyscore* zeigen auf, dass wo immer möglich der Full-Take-Ansatz praktiziert wird, wonach komplette Datenströme der weltweiten Kommunikation abgefangen und maschinell ausgewertet werden. Hierfür werden verdachtsunabhängig und auf globaler Ebene Daten direkt von Endgeräten, von Content-, Zugangsprovider, oder roh aus Funkverbindungen, Glasfaserleitungen und Unterseekabel ausgeleitet, erfasst, gegebenenfalls auf einer gemeinsamen Plattform gelagert und von dort aus abgerufen sowie verwertet.

Um die Jahrtausendwende wurde das *Echelon*-System der *Five Eyes*-Staaten (Australien, Kanada, Neuseeland, UK, USA) offiziell im Rahmen einer Untersuchung der Europäischen Union EU bestätigt.[4] Dieser Verbund ermöglicht schon seit Jahrzehnten die umfassende Überwachung von (zunächst: satellitenbasierter; später: generell funk- und kabelbasierter) Kommunikation.

Auch die Schweiz mischt im globalen Datenhandel mit: An drei Standorten überwacht die Eidgenossenschaft seit 2005 im Vollbetrieb mittels eines *Onyx* genannten Systems zur

---

[2] Aktivistin und Sprecherin des Chaos Computer Club CCC

[3] Tatsächlich lagen auch schon vor Jahrzehnten deutlich Hinweise vor, dass eine globale Überwachungsmaschinerie existiert: zu dieser Zeit allerdings in summarischer Erwähnung und wenig konkret. Zudem: noch nicht so fortgeschritten wie in den jüngeren Jahren.

[4] Erstmals wurde es schon 1988 enthüllt; vgl. Abschnitt 2.4.2.

Massenüberwachung von Funkverbindungen[5] satellitengestützte Kommunikationsströme in umfassender und (öffentlich) unkontrollierbarer Weise.

Während der Masterarbeit hat das Schweizer Parlament (National- und Ständerat) am 25. September 2015 zudem das Nachrichtendienstgesetz NDG verabschiedet:[6] Dieses sieht nebst vielen weiteren Überwachungsmöglichkeiten auch – ähnlich wie bei *Tempora* – vor, über Kabel geführte Kommunikation der Möglichkeit der Massenüberwachung zu unterstellen. Dafür sollen Schweizer Glasfaserleitungen direkt unter der zwangsweisen Datenzuleitung durch Schweizer Internet-Zugangsprovider angezapft werden können. Hinzu kommt, dass jüngere Medienberichte unter Beizug konkreter geheimdienstlicher Unterlagen darauf hinweisen, dass Schweizer Leitungen der *Swisscom* bereits heute von zumindest ausländischen Diensten angezapft werden. [38, 67]

Spätestens mit dieser in- und ausländischen Überwachung kabelgebundener Kommunikation muss jede in der Schweiz ansässige Person damit rechnen, von der Massenüberwachung betroffen zu sein, unabhängig davon, ob diese eine Gefahr für die innere oder äussere Sicherheit der Schweiz oder die eines anderen Landes darstellt.

Es ist öffentlich – wenn auch wenig konkret – bekannt, dass die überwachten Datenströme mittels sogenannter Selektoren oder (kombinierten) Suchbegriffen "gerastert"[7] werden, um die maschinell zu verarbeitenden Inhalte zu "filtern" und zu sortieren. In der Schweiz hat die für die Geheimdienstkontrolle zuständige Geschäftsprüfungsdelegation aus National- und Ständeräten GPDel diverse Berichte verfasst, wo das etwaige Vorgehen beschrieben wird, ohne aber je auf die genaue Natur des Zustandekommens der Selektoren einzugehen. Dennoch: die GPDel befasst sich kritisch mit der Wirksamkeit der angewandten Methoden und wirft Fragen der Vereinbarkeit der existierenden Praxis mit der Europäischen Menschenrechtskonvention EMRK auf.

Die Masterarbeit hat im Kern den Sinn, den Einsatz der Computerlinguistik und Sprachtechnologie für den Bereich der Massenüberwachung in ihrer Methodologie – soweit öffentlich ermittelbar – zu erörtern, in einem eigenen Überwachungssetting zu illustrieren und damit auch kritisch zu hinterfragen.

Immer wieder habe ich im Nebenfach der Soziologie die Möglichkeit genutzt, mich mit Themen der informationellen Selbstbestimmung, Freiheit und Demokratie im und mittels Internet auseinander zu setzen. Im Rahmen dieser Abschlussarbeit wird dieses Themenfeld

[5] Dem Wesen nach *Echelon*-ähnlich.

[6] Vgl. Geschäftsdatenbank Curia Vista (2015). 14.022 – Geschäft des Bundesrates. Nachrichtendienstgesetz. URL http://www.parlament.ch/d/suche/seiten/geschaefte.aspx?gesch_id=20140022. Abruf: 14. November 2015.

[7] Angespielt wird hier auf den Begriff der *Rasterfahndung*, die im Deutschland der 1970er Jahre eingeführt wurde, um nach RAF-Mitgliedern mittels massenweiser Auswertung von Datenbeständen unter Anwendung von Verfahren mit Ausschlusskriterien zu fahnden.

erstmals mit meinem Hauptfach der Computerlinguistik und Sprachtechnologie verknüpft: der Fokus der Masterarbeit ist naturgemäss ein computerlinguistisch- und technischer, wobei deren Befunde durchaus das Potenzial haben, mancherlei soziologische Fragestellung aufzuwerfen, die gesonders erörtert werden kann und – so denke ich – auch muss.

In vier Punkten liegt dieser Arbeit folgende Motivation zu Grunde:

1. Der global existierende und massive Überwachungskomplex, der mithin durch Entwicklungen im Bereich der Computertechnik und -linguistik begünstigt wird, soll vergegenwärtigt werden: dies geschieht durch Aufzeigen des Forschungsstandes im Bereich der Überwachungsmethodologie, insbesondere wo sie sich Techniken aus der Wissenschaft der Computerlinguistik oder Softwareprodukten mit sprachtechnologischen Fähigkeiten bedient; weiterhin durch Zitation offizieller Berichte parlamentarischer Kontrollgremien und mit Bezug zu nicht offiziellen Publikationen, wie solchen, die zunehmend durch Enthüllungen des ehemaligen Geheimdienstmitarbeiters Edward Snowden an die Öffentlichkeit geraten sind oder jenen, die als Veröffentlichungen seitens der Whistleblower- und Journalisten-Plattform *WikiLeaks*[8] publiziert werden. Damit soll das bestehende "Herrschaftswissen"[9] in Sachen Theorie und Praxis von Formen der Massenüberwachung offengelegt und dadurch entzaubert und entmystifiziert werden. Es soll zwar nicht darüber hinweggetäuscht werden, dass der Umfang der bestehenden und global existierenden Überwachung harmlos ist oder zu unterschätzen wäre, doch aber gezeigt werden, dass im Grundsatz die bekannt gewordene Praxis dem *Stand der Technik*[10] entspricht. Gegebenenfalls kann damit – mit Überraschung – festgestellt werden, dass das komplette "Arsenal" computerlinguistischen Wissens ausgeschöpft wird.

2. In der Schweiz finden politische Entwicklungen hin zu mehr und umfassenden Formen der Massenüberwachung statt, die methodisch computerlinguistisch abgestützt sein können: das Überwachungsgesetz BÜPF[11] für den (repressiven) Bereich der Strafverfolgung ist *einerseits* in Arbeit total revidiert zu werden, *andererseits* wurde kürzlich das Geheimdienstgesetz NDG[12] für den (präventiven) Geheimdienstbereich geschaffen, wogegen aus linken und netzpolitischen Kreisen das bis Januar 2016 laufende fakultative Referendum ergriffen wurde. Diese Gesetze kennen Formen der Massenüberwachung, die als *Antennensuchlauf*, *Funk-*, *Kabelaufklärung* oder *Vorratsdatenspeicherung* in der Schweizer Öffentlichkeit bekannt sind. Diese Arbeit soll für den bisher in der Schweiz wenig beachteten Teil der *Funk-* und *Kabelaufklärung* einen substanziellen Beitrag im Verständnis der möglichen Funktionsweise dieser

[8] Herausgeber: Julian Assange

[9] Damit ist im soziologischen Sinne gemeint: Wissen, das nur der regierenden und damit herrschenden Elite bekannt ist und ihr einen Informationsvorsprung gegenüber anderen Eliten einerseits und der Bevölkerung andererseits verschafft.

[10] Oder: *State-of-the-Art*

[11] Bundesgesetz zur Überwachung des Post- und Fernmeldeverkehrs

[12] Bundesgesetz über den Nachrichtendienst (auch: Nachrichtendienstgesetz)

Form der Massenüberwachung liefern, die auch einem breiten Publikum auf Basis dieser Arbeit zugänglich gemacht werden kann.

3. Diese Abschlussarbeit soll neben parlamentarischen Berichten und anderer Untersuchungsarbeiten, die weltweit im Bereich der Aufklärung der Massenüberwachung laufen, dazu anregen, auf technischer Grundlage eine kritische Diskussion über die Wirksamkeit von Massenüberwachung zu ermöglichen, die angeblich dafür betrieben wird und unabdingbar dafür da sein soll, die Sicherheit aller zu erhöhen – wofür aber auch alle und faktisch das Grundrecht auf Privatsphäre aufgeben sollen.

4. Zu guter Letzt soll die Arbeit in motivationaler Hinsicht dafür sensibilisieren, das politische Missbrauchspotenzial der Wissenschaft der Computerlinguistik und ganz besonders ihrer sprachtechnologischen Anwendungen aufzuzeigen: ich sehe dies als die Aufgabe der kritischen Reflexion des eigenen Fachs, die im Sinne von Verantwortungsbewusstsein und (sozialer) Nachhaltigkeit nach meinem Empfinden zu den Kernkompetenzen aller Personen gehören, die entsprechende Kenntnisse erwerben und das Privileg haben, in höheren sozialen Stellungen zu sein, die mit erheblichem – wenn auch nur potenziellem – Einfluss auf die Gesellschaft und ihren nachfolgenden Generationen verbunden sind. Dies steht auch ganz im Einklang mit dem Leitbild der Universität Zürich, wonach *“[z]u verantwortlicher Wissenschaft [...] die ethische Reflexion ihrer Mittel und Folgen für Mensch, Tier und Umwelt [gehört]”*. Diesem Grundsatz sei diese Arbeit verpflichtet. [121, S.10]

1.2 Forschungsfragen und Hypothesen

Im Grundsatz leiten die folgenden Forschungsfragen diese Arbeit an:

- Was ist Überwachung, spezifisch Massenüberwachung?
- Was sind (mögliche) *computerlinguistische* und damit theoretische Grundlagen, die für Massenüberwachung geeignet sind und wie kann Massenüberwachung *sprachtechnologisch* und damit praktisch umgesetzt werden?
- Was wurde im Kontext der Computerlinguistik für Zwecke der Massenüberwachung bisher erforscht: im Rahmen der NSA oder von Forschungsprogrammen der EU?
- Welche Formen der Massenüberwachung finden sowohl global als auch lokal (in der Schweiz) Anwendung und welche davon haben eine (starke) computerlinguistische Grundlage?
- Welche konkreten *sprachtechnologischen* Produkte sind für die Massenüberwachung vorhanden?
- Welche linguistisch verwertbaren Daten fallen bei der Überwachung vollständiger Datenströme überhaupt an?

- Welche konkreten Daten eines Datenstroms sind am interessantesten und reichen aus, um die Natur der Massenüberwachung im Rahmen der Masterarbeit exemplarisch darzustellen?

Diese Fragenkomplexe sollen im Rahmen des Rechercheteils der Arbeit im Allgemeinen beantwortet werden.

Zusätzlich führt die Masterarbeit im Kern ein kritisch-illustratives Datenprojekt, das aufzeigen soll, wie auf Basis von inhaltlichen Selektoren (auch: Soft-Selektoren, Suchbegriffen, Stichworten oder Keywords) zur Massenüberwachung mehrheitlich Falschtreffer entstehen, womit Massenüberwachung – extrapoliert auf weitere Gebiete, wo ähnliches Vorgehen betrieben wird – zu vielen Falschverdächtigungen führen kann.

Drei Hypothesen werden im Rahmen dieses Teils der Arbeit als wahr angenommen und sollen durch eigene Versuche plausibilisiert werden:

1. Werden allgemeine Datenströme oder Dokumentenkollektionen mit spezifischen Selektoren inhaltlicher Art – auf Basis eines gegebenen Verdachtskorpus – durchsucht, ist der Anteil der False-Positives prävalent.[13]
2. Der Anteil der False-Positives variiert unter den Bedingungen der ersten Hypothese je nach Güte und Parameter des verwendeten computerlinguistischen Modells, Selektoren zu generieren, merklich: dennoch dominieren auch bei "besseren"[14] CL-Modellen False-Positives die Ergebnisse.
3. Selektoren von allgemeinsprachlicher Bedeutung, die – wenn sie auch daraus generiert wurden – nicht spezifisch auf das gegebene Verdachtsmaterial angepasst scheinen, erzeugen zwar eine höhere Trefferquote, doch sind damit auch mehr False-Positives verbunden.[15]

1.3 Fokus und Grenzen

Der Arbeit hat den Fokus exemplarisch zu sein und weist – in Selbstkritik betrachtet – Einschränkungen auf. Sowohl mit der Wahl der Daten als auch Methoden sind der Arbeit entsprechend Grenzen der Aussagekraft und der Repräsentativität gesetzt, die zumindest

[13] Das bedeutet übertragen, dass bei einer Massenüberwachung, bei der unterschiedslos aller verfügbarer Textinhalt auf Basis einer kategorienspezifischen Stichwortsuche "gerastert" wird, die weitaus meisten Treffer, die als Filtrat bleiben und als "verdächtigen" Text im Sinne der Suche gelten, False-Positives sind.

[14] Es muss betont werden, dass jedes Modell durch seinen jeweiligen Fokus auf den Text, den es in Form von Selektoren nachbilden soll, andere – möglicherweise ausschliessende – Selektoren generiert, so dass selbst bei vergleichbaren Anteilen der False-Positives bei unterschiedlichen Selektoren veschiedener Modelle, gänzlich andere Inhalte (falsch) filtriert sein können.

[15] Das bedeutet, dass wenn auch ein Korpus mit "extremistischen" Inhalten dafür gewählt wird, Selektoren abzuleiten, diese aus Wörtern oder Ausdrücken konstituiert sein können, die soziolinguistisch wenig spezifisch auf die repräsentierte Gruppe ausfallen können.

die Folgenden sind:

- Im Rechercheteil von Kapitel 2 wird (1) nur ein Teil der (Forschungs-)Literatur aufgezeigt, die sich mit der Massenüberwachung unter Einsatz computerlinguistischer Mittel beschäftigt; (2) die referenzierte Literatur kann nur in Auszügen und partiell kritisch analysiert werden, so dass das Bild vom Stand der Technik nicht vollständig sein kann.
- Der Umfang der durchgeführten Volldatenüberwachung des eigenen Internetanschlusses zur Generierung realistischer Evaluationsdaten beschränkt sich auf einen konkreten Zeitraum von zehn Tagen, womit der Datenbestand (1) im Sinne des Sparse-Data-Problems zu knapp ausfallen mag: in realen Szenarien ist möglich, dass eine Erfassung über einen längeren Zeitraum erfolgt; (2) ist die Repräsentativität des Datenstroms nicht gegeben, weil ein Bias in sowohl zeitlicher als auch persönlicher Hinsicht besteht.
- Der eigens generierte Datenstrom stellt nur eine von vielen möglichen Perspektiven dar, mit der Einsicht in private Kommunikationen genommen werden kann: Potente Geheimdienste verfügen über eine globalere Sicht von Kommunikationen und können beispielsweise Datenströme zwischen Servern[16] mitschneiden oder durch Einsatz sogenannter GovWare (oder: Staatstrojanern) auch Daten auf Endgeräten einsehen.[17]
- In den Fällen, wo als Evaluationssysteme Suchmaschinen zum Einsatz kommen, muss auf die Ergebnisse der entsprechenden Anbieter vertraut werden: es kann nicht ausgeschlossen werden, dass die resultierenden Treffer anderweitig "gefiltert" (bei den grösseren Anbietern mithin: zensiert) sind. Dieses Problem stellt sich bei dem eigens generierten Evaluationskorpus im Gegenzug nicht.
- Durch die Auswahl der konkreten Trainingsdaten wird ein Raum der überhaupt möglichen Selektoren aufgespannt, der naturgemäss durch die Anzahl der verfügbaren Types[18] beschränkt ist: es gibt Hinweise dafür, dass in der real existierenden Praxis der Massenüberwachung Selektoren ebenfalls manuell festgelegt werden können. Dies wird im Rahmen dieser Arbeit missachtet.
- Die Trainingsdaten sollen nicht "Links-" und "Rechtsextremismus" repräsentieren, sondern zwei konkrete Gruppen, die diesen Spektren zuzuordnen sind; auf Grund des Fokus auf zwei konkrete Webseiten können diese Korpora entsprechend den Anspruch der Repräsentativität nicht erfüllen, für ein jeweils komplettes Spektrum zu stehen. Sehr wohl aber sind sie geeignet, zu testen, ob es damit gelingt weitere ähnliche Gruppen zu finden. Ganz allgemein gefasst: Es wird ein nur eingeschränkter Bereich

[16]Dies ist zum Beispiel zum Mitlesen von E-Mails relevant.

[17]Diese Einschränkung bedeutet im Gegensatz zur vorangegangenen Einschränkung, dass unter Einsatz von Überwachungsmöglichkeiten, die durchdringlicher sind, selbst im Rahmen von einer Volldatenüberwachung von "nur" zehn Tagen mehr linguistisch verwertbare Daten hätten gesammelt werden können. Das Sparse-Data-Problem würde geringer ausfallen.

[18]Gemeint ist die Menge der vorhandenen Wortoberflächen.

möglicher Überwachungskategorien erfasst. Bereiche etwa der diplomatischen oder Wirtschaftsspionage, die nachweislich als weitere Überwachungskategorien existieren und im globalen Ausmass betrieben werden, sind hiermit nicht abgedeckt.

- Annahmen werden getroffen, wie Selektoren korpusbasiert (statistisch) und zudem linguistisch motiviert (durch Wortlisten zugeschriebener Bedeutung) zustande kommen müssten; es können auch andere Annahmen getroffen werden. Die drei Überwachungsmodelle stellen bloss Beispiele einfacher und komplexerer Verfahren zur Generierung von Selektoren dar: es sind nebst der Wahl anderer Modelle auch Abwandlungen der ausgewählten Modelle möglich.
- Es werden auf inhaltlicher Ebene keine Filterstufen verwendet oder mögliche Negativ-Selektoren genutzt, um prohibitive Begriffskombinationen auszuschliessen, wie dies beim BND im Rahmen der G10-Filterung geschehen sollte.[19]
- Eine Unterscheidung nach Metadaten und eigentlichen Inhaltsdaten wird nicht vorgenommen: alle Daten textueller Natur werden durchsucht. Entsprechend sind die resultierenden Treffer offene Ergebnisse, die nicht durch formale Suchbegriffe (wie IP-, E-Mail-Adressen oder Ähnlichem) beschränkt sind.

Die Arbeit erhebt trotz aller Einschränkungen dennoch den Anspruch, geeignet zu sein, die Natur eines Teils der globalen Massenüberwachung plausibel darzulegen, wenn auch die konkret resultierenden Evaluationsergebnisse von Kapitel 4 – entsprechend den oben genannten Foki und Grenzen der Arbeit – kritisch zu hinterfragen sind.

1.4 Aufbau

Die Arbeit zeigt im Recherche-Teil von Kapitel 2 auf, was in Sachen *Massenüberwachung* und spezifisch *computerlinguistisch* möglich ist. Quellenmaterial ist überwachungsspezifische Forschung und Möglichkeiten und Praxis heutiger Überwachung, die beispielsweise durch die Enthüllungen Edward Snowdens bekannt geworden ist.

Im praktischen und betont kritisch-illustrativen Teil von Kapitel 3 wird im Rahmen eines eigenen Überwachungssettings aufgezeigt, wie Massenüberwachung auf Basis von Suchbegriffen (auch: Selektoren) funktionieren könnte. Drei Modelle werden beschrieben und dazu genutzt, in der Summe 140 Selektoren automatisch zu generieren, die dazu eingesetzt werden könnten, "extremistische" Treffer links- und rechtspolitischer Art zu erzeugen.

Kapitel 4 widmet sich der manuellen Evaluation der drei Modelle: im Zuge dessen werden auch besondere True- oder False-Positives hervorgehoben, die durch zwei unabhängige Annotationen ermittelt wurden. Es wird damit der Versuch unternommen, die Natur der Massenüberwachung auf Basis von Selektoren aufzuzeigen.

[19] Vgl. Abschnitt 2.4.3.1

Im Schlussteil von Kapitel 5 werden zunächst die resultierenden Treffer in ihrer Evaluation kritisch diskutiert und weitere Fragen allgemeiner und weitergehender Art aufgeworfen; die eingangs gestellten Forschungsfragen und Hypothesen werden – soweit möglich – summarisch beantwortet, plausibilisiert und hinterfragt; schliesslich wird die Arbeit mit einem Fazit, das den Erkenntnisgewinn synthetisiert, zu einem Ende geführt.

2 Stand der Technik des Überwachungskomplexes

> *Fängt man etwas Interessantes mit seinem Leben an, sind die Informationen auf dem eigenen Computer irgendwie von Bedeutung oder geldwert für feindlich gesinnte Mitmenschen oder Entitäten, dann muss man damit rechnen, auch gezielt digital angegriffen zu werden. Sei es als politischer Aktivist, recherchierender Journalist, Wissenschaftler, Medienberühmtheit, provokanter Künstler, Manager oder Politiker – sobald man im Besitz von Informationen ist, die für jemand anderen von Interesse sind, kann man ins Fadenkreuz geraten. Die Spannbreite reicht von intimen Fotos bis zu Kundenlisten oder technischen Planungsdokumenten. Man muss sich klarmachen, dass sich in den letzten Jahren im Windschatten der finanziell potenten Geheimdienste ein professioneller Industriezweig herausgebildet hat, der das Hacken als einträgliches Geschäft betreibt.*
>
> – Constanze Kurz, Frank Rieger, Harald Staun und Sascha Lobo
> am 12. Juli 2015 [59]

Zunächst ist festzuhalten, dass die US-amerikanische National Security Agency *NSA* von der eigenen Regierung das Recht erhalten hat, nominell alle ausser vier Länder auszuspähen, welche sind: Australien, Grossbritannien, Kanada und Neuseeland. Diese sind mit den USA zusammen Teil des Five Eyes-Verbundes *FVEY*, der eine globale Abhörmaschinerie betreibt. [73]

Durch den Datenaustausch, den die *FVEY*-Staaten untereinander pflegen, kann über den Partnerumweg, der Datenverkehr der eigenen Bevölkerung dennoch überwacht werden, wie dies beispielsweise für das Vereinigte Königreich belegt ist. [76]

In aller Regel wird der Ausbau der Überwachung mit der Gefahr vor Terrorismus begründet. Gemäss National Intelligence Priorities Framework *NIPF* ist dieser Bereich "Counterterrorism" oder "TERR" allerdings nur eines von 32 Themenfeldern, die geheimdienstlich relevant sind. [15][96, S.149]

Weitere Überwachungsbereiche, die vermuten lassen, dass Geheimdienste nebst Sicherheitsvielmehr Machterhaltungs- und Wirtschaftsinstrument sind, eröffnen sich mit Beispielen der weiteren Bereiche, die bestehen:

- Leadership Intentions (LEAD): Regierungsabsichten

- Economic and Financial Stability (ECFS): Ökonomische und finanzielle Stabilität
- International Trade Policy (TRAD): Internationale Handelspolitik
- Food Products and Security (FOOD): Nahrungsmittelprodukte und -sicherheit
- Democratization: Demokratisierung
- Environment and Natural Ressources (ENVR): Umwelt und natürliche Ressourcen
- Demographics (DEMG): Demografie
- Health and Infectious Diseases (HLTH): Gesundheit und Infektionskrankheiten
- Cyber (CYBR): Cyberspace / Internet

In der geleakten "United States SIGINT System"-Liste von Januar 2007 werden weiterhin explizit Länder wie China, Deutschland, Frankreich, Indien, Israel, Japan, [Süd-]Korea[20], Russland oder Schweden geführt. Diese seien sogenannter Signals Intelligence *SIGINT* zu unterziehen, weil sie gemäss der Überwachungsmission "Emerging Strategic Technologies" das Potenzial hätten, kritische Technologien zu entwickeln, welche sie in strategischer Hinsicht militärisch, ökonomisch oder auch politisch – gegenüber den USA – bevorteilen könnten. [80]

Ein sehr deutlicher Hinweis, dass auch gewöhnliche Bürger konkretes Ziel geheimdienstlicher Aktivitäten sein können, bietet ein *The Intercept*-Artikel vom März 2014:

> *The internal post – titled "I hunt sys admins" – makes clear that terrorists aren't the only targets of such NSA attacks. Compromising a systems administrator, the operative notes, makes it easier to get to other targets of interest, including any "government official that happens to be using the network some admin takes care of."*

Demnach besteht ein internes Dokument über die "Jagd" auf Systemadministratoren, das deutlich macht, dass über solche Personen auch der Zugang zu den eigentlich interessierenden Zielen (im Beispiel: Behördenmitglieder) erfolgen kann, deren Netzwerk unter der Administration der konkret angegriffen Personen stehen. [28]

2.1 Zu Wesen und Spielarten der Massenüberwachung

2.1.1 Was ist Überwachung und spezifisch Massenüberwachung?

Joachim Scharloth schildert in seinem Blog [102] eine weithin bekannte Überwachungssituation, die durch ihre spürbare und unmittelbar Natur dazu beiträgt, dass sich normkonformes

[20] Es ist naheliegend, dass mit "Korea" Südkorea gemeint ist.

Verhalten einstellt – das Fliegen mit den bekannten Flughafenkontrollen:

> *Mit dem Flugzeug zu reisen hat bei allen Vorzügen einen entscheidenden Nachteil: Keine andere Form des Reisens normiert die Passagiere so weitreichend wie eine Flugreise. Sie erlaubt den Reisenden nur eine bestimmte Menge Gepäck in vorgeschriebener Form, weist ihnen einen engen Raum zu, den sie auch nur zu ganz bestimmten Zwecken verlassen dürfen, zwingt auf visuelle Signale hin zum Anschnallen, zwingt zum Ausschalten von Geräten und – indem das Entertainment-Programm unterbrochen wird – zum Zuhören bei allen Ansagen. Und keine andere Form des Reisens kennt derlei Sanktionen, wenn man sich der Normierung widersetzt: abhängig vom Land können einem Raucher auf der Bordtoilette Strafen vom Bussgeld bis zur mehrmonatigen Gefängnisstrafe blühen. Die Annehmlichkeit der schnellen Überbrückung von Entfernungen zu einem noch erträglichen Preis wird also durch die Akzeptanz einer weitgehenden Normierung erkauft.*

Dieses treffende Beispiel erlaubt es, gut nachzuvollziehen, dass Überwachung – wenn sie spürbar ist – zu weniger abweichendem Verhalten führt.

Entsprechend lässt sich das Gefühl von Überwachung auf dieser Basis nach Spürbarkeit in zumindest folgenden Dimensionen besser fassen:

- Spürbarkeit der Überwachung auf Grund der Natur des Überwachers (Mensch versus Maschine)
- Spürbarkeit der Überwachung auf Grund körperlicher Nähe des Überwachers (Polizist im physischen versus Überwachungssoftware im virtuellen Raum)
- (Gefühlter) Grad der Betroffenheit (Überwachung aller versus Überwachung Einzelner)

Bezogen auf das Internet, insbesondere den da vorhandenen Möglichkeiten Kommunikation massenweise, unspürbar abzufangen und automatisiert auszuwerten, wird – werden diese drei Dimensionen zur Operationalisierung der Spürbar- oder Fühlbarkeit überwacht zu sein – beigezogen, rasch klar, dass auf Grund

1. der maschinellen Natur der Überwachung im Netz;
2. der nicht vorhandenen Körperlichkeit mangels der physischen Spürbarkeit der im Netz laufenden Überwachung und
3. mangels der Erkennbarkeit überhaupt überwacht zu werden sowie der Beschwichtigungen seitens der politischen Führung, bloss "Terroristen" würden gesucht,

sich das Gefühl einstellen kann, eine Überwachung existiere – spürbar – nicht.

Hinzu weist Scharloth auf den Begriff des "Filterns" hin, welchem oft die (politische)

Funktion zukommt, die Massenüberwachung zu verharmlosen:

> *Gerne greift die Überwachungsapologetik auf den Begriff des "Filterns" zurück, um zu implizieren, dass nur ein geringer Teil des Materials angeschaut wird: unter Generalverdacht am Anfang aber steht alle Kommunikation. Eine oberflächliche Analyse, welche nicht den Anspruch erhebt, den Gesamtsinn einer Kommunikation zu erfassen, sorgt dafür, dass Material (zur weitergehenden automatischen oder manuellen Analyse) zurückgehalten wird.*

Tatsächlich aber findet Überwachung nicht erst statt, wenn ein Text "gefiltert" wurde, sondern schon vorher – bei der Erfassung; hierbei verweist Scharloth auf ein Urteil des deutschen Bundesverwaltungsgericht von 1999, das zu ähnlichen Schlüssen kommt, wenn das Gericht es auch für haltbar hält, Massenüberwachung unter bestimmten Bedingungen zu betreiben.

Massenüberwachung bezeichnet also eine Überwachungsform, die nicht zielgerichtet, sondern verdachtsunabhängig und anlasslos erfolgt: sie spekuliert darauf, Verdächtigungen erst zu schaffen oder einen Anlass zur Verdachtsschöpfung erst noch zu begründen.

In Reinform kann die uns interessierende digitale Massenüberwachung im Internet zudem nach zwei Foki und entsprechend zwei Methodologien erfolgen:

- Inhaltsdaten: Es interessieren die Inhaltsdaten einer Kommunikation; der Verdacht soll auf Grund des Inhalts eines Kommunikationsvorgangs konstituiert werden. Inhalte können beispielsweise Chat-Nachrichten, E-Mails, Webseiten-Inhalte oder heruntergeladene PDF-Dateien sein.

- Metadaten: Es interessieren die Metadaten der Kommunikation, das heisst der Verdacht soll sich nicht aus den Gegenständen selber, sondern dem Umstand deren Verbindung herleiten. Verbindungsdaten können zum Beispiel Angaben zu Kommunikationspartner, aufgerufene Webseiten-URLs oder Telefon-Nummern sein.

Im Rahmen dieser Arbeit interessiert die Form der Massenüberwachung, die auf Inhaltsdaten abzielt. Eine Arbeit, die sich kritisch mit der Massenauswertung von Metadaten beschäftigt, hat Daniel Mossbrucker kürzlich vorgelegt: sie geht insbesondere der Frage nach, inwiefern der Quellenschutz bei journalistischen Recherchen noch gewahrt werden kann, wenn ein Land alle Metadaten aller Kommunikationsteilnehmer speichert. Er zeigt auf, dass die Verfügbarkeit von Informationen über die "blossen" Umstände von Kommunikation bereits ausreicht, den journalistischen Quellenschutz auszuhebeln. Dies gilt selbst dann, wenn die gespeicherten Metadaten für nur wenige Wochen zugänglich sind. [72]

2.1.2 Welche Ansatzpunkte zur digitalen Überwachung gibt es?

> *Warum können wir nicht alle Signale sammeln, jederzeit?*
>
> – Keith Brian Alexander[21], 2008 [96, S.120]

In nicht abschliessender Weise werden hier Angriffsvektoren aufgezeigt, wie sie aus dem Gebiet der IT-Sicherheit entstammen und die eingesetzt werden können, um an Daten heranzukommen, welche in der Folge linguistisch und maschinell verwertet werden können.

Wie im Abschnitt 2.1.2 dargelegt wird, bedienen sich insbesondere global operierende Geheimdienste potenter Staaten aller nur denkbaren Angriffsvektoren und übertreffen in der Breite und Tiefe ihrer Möglichkeiten bei weitem kriminell organisierte Gruppen oder aktivistische Hackerkreise, die aus politischen Motiven agieren.

2.1.2.1 Überwachung von Datenströmen

In "klassischer" Hinsicht lassen sich Datenströme abfangen, speichern, scannen und auswerten, wo Computer oder Peripheriegeräte[22] über Funk- oder Kabelverbindungen miteinander zu einem Netzwerk verknüpft sind und Daten miteinander austauschen.

Ohne auf Spezifika häuslicher, institutioneller, nationaler oder globaler Netzwerke einzugehen, sei abstrakt festgehalten, dass Signale zwischen Computern an folgenden Stellen abgefangen werden können:

(a) Auf lokaler Ebene zuhause, Institutionen, im Staat oder auch im Rahmen lokal anmutender, physisch allerdings über unsichere Verbindungen aufgespannter Virtuelle Private Netzwerke (VPN), die von aussen als verschlüsselte Datenströme erkannt werden.[23]

(b) Auf Ebene der Zugangsprovider, welche Privaten, Institutionen oder Staaten den Zugang über Kabel- oder Funkverbindungen zum Internet ermöglichen.

(c) Auf nationaler oder globaler Ebene zwischen Providern[24] aller Art.

(d) Auf globaler Ebene bei Überwachung von transnationalen Glasfaserleitungen, Unterseekabeln oder Satellitenverbindungen oder auch sogenannten Internet Exchange Points, welche eigentliche Internetschaltstellen zwischen Content Delivery Networks CDNs und Internetprovider darstellen.

[21]Direktor der NSA von 2005–2014

[22]Das können Mäuse, Tastaturen, Drucker und andere Geräte sein, die (als peers) an anderen Endgeräten oder (als Clients) an Serversystemen angeschlossen sind.

[23]Zumindest von Innen ist Überwachung und Kontrolle des Netzwerkes seitens der VPN-kontrollierenden Instanz möglich: es besteht die Sicht vergleichbar jener eines Zugangsproviders.

[24]Damit sind auch Content-Provider wie E-Mail-Dienstanbieter, Social-Media-Plattformen usw. gemeint.

2.1.2.2 Ausleitung Daten Endgerät (an beiden Enden)

Unter der Bedingung, dass entweder Soft- oder Hardware von Anfang an unterwandert beziehungsweise verwanzt ist oder sich durch Sicherheitslücken zumindest unter Kontrolle bringen lässt, können die interessierenden Daten für eine Überwachung direkt an der Datenquelle beziehungsweise dem unmittelbaren Datenziel ausgeleitet werden.

Endgeräte wie Notebooks, Smartphones, Smart-TV-Geräte oder selbst – nicht als vernetzt intendierte Geräte wie – Wasserkocher[25] können von Anfang an mit Malware der Unterart Spyware ausgestattet sein, und den Internetverkehr insgesamt oder in den gewünschten Datensätzen an interessierende Dritte auszuleiten.

In Fällen, wo ein Endgerät noch nicht infiziert ist, können Methoden wie Deep-Packet-Injection eingesetzt werden, um im Rahmen eines eingehenden Verkehrs ein Gerät zu infiltrieren, indem der vom Gerät empfangene Datenstrom durch Malware ergänzt wird – in einer Form, dass auf dem Zielsystem eine Sicherheitslücke für die Installation derselben ausgenutzt wird.[26] Ist der End-User die eigentliche Schwachstelle im zu überwachenden Komplex, können "klassische" Angriffsverfahren wie der Versand von syteminfiltrierenden Mailanhängen oder Drive-By-Attacken eingesetzt werden, wo dem Opfer ein Anreiz[27] gegeben wird, einen Anhang zu öffnen oder auf einen Link zu klicken.[28]

2.1.2.3 Ausleitung Daten Servergerät (bei Client-Server-Modellen)

Ebenfalls zu beachten ist die Möglichkeit der massenweisen Ausleitung von Daten von Serversystemen (gerade von grossen Anbietern) wie das von Edward Snowden enthüllte *PRISM*-Programm der *FVEY* [96, S.127] nahelegt. Hierbei wird ein Datenabfluss bei Content-Anbietern organisiert: Sofern die Daten für Anbieter wie *Google* oder *Facebook* selber lesbar sind – was der Regelfall ist –, so muss es nicht einmal einen Unterschied machen, ob CA-basierte Verschlüsselung wie TLS (für HTTPS) benutzt wird, weil der Angriffspunkt beim Anbieter direkt und nicht bei der Transportverbindung zum Benutzer hin liegt.

[25] Es gibt dokumentierte Fälle von Haushaltsgeräten, die (unnötigerweise) mit WLAN-Technologie ausgestattet sind, um automatisiert Angriffe auf ansprechbare Access Points auszuführen und (nach Möglichkeit) Verkehr erfolgreich angegriffener Geräte auszuleiten. Vgl. [110].

[26] Solche "Injektionen" können beispielsweise im Rahmen eines Systemupdates oder beim Herunterladen einer Datei erfolgen. Der Angriff erfolgt in der Regel völlig unbemerkt und löst beim End-User keinen Verdacht aus.

[27] Es können beispielsweise öffentliche Profilmeldungen auf Social-Media-Plattformen analysiert werden, um plausible Nachrichten an einen Benutzer zu präparieren.

[28] Durch die Erforderlichkeit der User-Interaktion handelt es sich im letzten Fall um einen Angriff der Art des Social-Engineering

2.1.3 Beispiele: Fälle und Folgen von Verdächtigungen auf linguistischer Basis

> *Donnez-moi deux lignes de la main d'un homme, et j'y trouverai de quoi suffire à sa condamnation.*
>
> – Armand Jean du Plessis de Richelieu[29]

In diesem Abschnitt werden drei Fälle geschildert, wo Personen auf Grund linguistischer Merkmale in Texten, die sie verfasst haben (sollen), für Taten verdächtigt oder gar verurteilt wurden.

2.1.3.1 Schweiz: "Rütli-Bomber" ungrammatikalisch in seiner Rache

Am 1. Januar 2007 detonierte an der Jahresfeier der Schweizerischen Eidgenossenschaft auf der Rütli-Wiese ein selbstgebauter Klein-Sprengsatz: niemand wurde verletzt. Gleichzeitig wurden im unmittelbaren zeitlichen Umfeld diverse Briefkästen von Mitgliedern der "Rütli-Kommission" gesprengt. Auf Grund eines anonymen Hinweises gegenüber der Kantonspolizei Aargau und einer Gefahreneinstufung sowie Informationen seitens des damaligen Schweizer Inlandsgeheimdienstes DAP kam es zu einer massiven Überwachung einer Person, die landesweit als "Rütli-Bomber" bekannt wurde. Die Überwachungen liefen trotz ihres invasiven Charakters[30] ins Leere. Schliesslich wollte die Schweizer Bundesanwaltschaft BA die Quelle der ursprünglichen Information darüber, dass es sich bei der als "Rütli-Bomber" verdächtigten Person tatsächlich um die gesuchte Person handelt. Sie stiess beim Geheimdienst auf taube Ohren. Die BA wandte sich an die Schweizer Regierung: diese fiel den (politischen) Entscheid, die Quelle sei zu schützen. Damit musste die BA das Verfahren mangels auch nur stichhaltiger Beweise einstellen.

Dem fälschlich Verdächtigten wurde das Leben offensichtlich zerstört: er hat den Kontakt zu Familie, Kind und Freunde verloren. In der Folge soll er sich bei Personen gerächt haben, die für seine Lebenssituation verantwortlich sein könnten und wurde dafür belangt.

Auf Grund von Drohbriefen, die er formuliert haben soll, verurteilte [116] ihn die Staatsanwalt auf Basis grammatikalischer und syntaktischer Merkmale, die – linguistisch betrachtet – einen wenig spezifischen Eindruck vermitteln:

> *Der zuständige Staatsanwalt sah den nicht geständigen Elektromonteur als überführt an. Auch wegen den grammatikalischen Mängeln des Beschuldigten, der in seinen Schreiben immer wieder durch die gleichen Kommafehler und einen notorisch falsch gewählten Akkusativ aufgefallen war.*

[29]Zumindest wird ihm dieses Zitat zugeschrieben; vgl. französischsprachige Wikipedia (2015).
Online: https://fr.wikipedia.org/w/index.php?title=Armand_Jean_du_Plessis_de_Richelieu&oldid=120075282#Citations_et_maximes_c.C3.A9l.C3.A8bres
(Abruf: 3. November 2015)

[30]Zum Beispiel wurde ohne gesetzliche Grundlage ein "Staatstrojaner" eingesetzt.

Der “Rütlibomber” selber beteuerte seine Unschuld.

2.1.3.2 Deutschland: Andrej Holm als Soziologe zum “Terroristen”

Bei Andrej Holm handelt es sich um einen Soziologen, der unter anderem zu Phänomenen der Stadtaufwertung (in der Soziologie auch: Gentrification) forscht. 2006 wurden gegen ihn und weitere Verdächtige Ermittlungen wegen “Mitgliedschaft in einer terroristischen Vereinigung” nach §129 des deutschen Strafgesetzbuches StGB aufgenommen: im Fokus standen Untersuchungen zum Zusammenhang “militante gruppe” oder “mg”, die gemäss eigenen Mitteilungen für mehrere Brandanschläge im Grossraum Berlin und Brandenburg verantwortlich zeichnete. Die Ermittlungen führten nach einer rund einjährigen invasiven Überwachung zur isolativen Untersuchungshaft. Als ein wichtiges Verdachtsmoment führte die Bundesanwaltschaft Begriffsverwendungen in Texten sowohl der “mg” als auch solchen (wissenschaftlichen) von Andrej Holm an, von denen er viele offenbar – öffentlich auffindbar – publizierte.

Der Verdacht wurde mittels *Google*-Suche [118] auf Basis von Einzelworten oder Ausdrücken begründet, von denen folgende öffentlich dokumentiert [26, S.69] sind:

(1) Nominal: `gentrification`, `prekarisierung`, `reproduktion`

(2) Verbal: `implodieren`

(3) Phrasal: `politische praxis`

(4) Adjektivistisch: `drakonisch`, `marxistisch-leninistisch`

Weitere Verdachtsmomente (ausser linguistische oder solche der “intellektuellen Urheberschaft”) waren zwar auch vorhanden, allerdings sind diese gemäss einem Offenen Brief mit Unterschrift internationaler Forscherkreisen zur Unterstützung von Andrej Holm genauso fragwürdiger Basis: so wurde eine “Kontaktschuld” wegen Freundschaftsbeziehungen oder Adressbucheinträgen konstruiert oder “konspiratives Verhalten” festgestellt, weil Andrej Holm und andere Verdächtige sich teilweise ohne Mobilgeräte trafen. [25]

Unter dem internationalen Protest wurde Andrej Holm nach wenigen Wochen aus der Untersuchungshaft entlassen, doch erst 2010 für unschuldig befunden. Die massive Überwachung lief in den Jahren dazwischen (offen) weiter, wie seine Lebenspartnerin Anne Roth in einem Blogpost zur “Innenansicht einer Terrorismus-Ermittlung” darlegt. [97]

2.1.3.3 USA: "Unabomber" Theodore Kaczynski und seine Phrase

Theodore Kaczynski (auch: "Unabomber"[31]) ist ein grün-anarchistischer Mathematiker, der die USA über Jahrzehnte mit einer landesweiten Briefbombenkampagne auf Trab hielt. Sein Anliegen war, die technologische Entwicklung dadurch aufzuhalten, dass die Exponenten derselben angegriffen werden. In einen längeren Text legte er seine Gründe dafür dar. Dieser Text wurde ihm aber zum Verhängnis, weil es dem FBI gelang, durch einen Tipp seitens David Kaczinskys[32] und Methoden aus der forensischen Linguistik der Autorenidentifikation, auf ihn zu schliessen.

Das US-amerikanische National Museum of Crime & Punishment [85] erwähnt eine bestimmte Phrase, die dem "Unabomber" eigen gewesen sein soll:

> *[T]he serial bomber sent a very long manifesto called Industrial Society and its Future to several publications demanding it be published. When they obeyed, a man named David Kaczynski read the manifesto and found it disturbingly familiar; the word choices and philosophy resembled those of his brother Theodore. There were particular phrases David recognized as Ted's, including a reversal of the common saying "have your cake and eat it too;" Ted preferred to say "eat your cake and have it too." These were unique enough to be instantly recognizable, but were not the only indicators.*

Andererseits wurden auch andere linguistische Merkmale – wie Wortfrequenzen – beigezogen, um die Verdächtigung zu plausibilisieren, die erforderlich war, um mittels richterlichen Beschluss einen Zugriff auf Theodore Kaczinsky vollziehen zu können. Dieser bestätigte seine Urheberschaft.

2.2 Computerlinguistische Forschung zur Ausübung von Massenüberwachung

Diese Sektion fokussiert auf Forschung, die im Bereich der Computerlinguistik durchgeführt wurde, mit dem offenen Zweck deren Ergebnisse für Zwecke der Massenüberwachung einzusetzen. Sie geht eingangs auf die Rolle der Computerlinguistik hierfür ein.

2.2.1 Rolle der Computerlinguistik für die Massenüberwachung

Bei der Computerlinguistik handelt es sich um ein interdisziplinäres Forschungsfeld, das sich zwischen (allgemeiner) Sprachwissenschaft und Informatik bewegt. Gemäss dem

[31] Bezeichnung im Rahmen der FBI-Operation *UNABOMB*: University and Airport Bomber, weil entsprechende Vertreter Ziel von Anschlägen waren.

[32] Bruder von Theodore Kaczinsky

Institut für Computerlinguistik der Universität Zürich untersucht Computerlinguistik

- *"wie die menschliche Sprache als Mittel zur Übermittlung, Speicherung und Verarbeitung von Information verwendet wird, und*
- *wie man diese Prozesse auf dem Computer modellieren und für konkrete Anwendungen nutzbar machen kann."*

Zudem wird demnach von Sprachtechnologie gesprochen, *"[w]enn die praktische Entwicklung von Werkzeugen interessiert, die in verschiedenen Phasen der automatischen Verarbeitung natürlicher Sprachen eingesetzt werden [...]"*. [122]

Weder der Computerlinguistik noch Sprachtechnologie ist Massenüberwachung inhärent: allerdings reichen ihre Ursprünge auf die Anfänge der Computerisierung zurück, wo insbesondere die US-Amerikaner russische Botschaften aus Forschung und Militär ins Englische zu übersetzen versuchten. Am Anfang taten sie das, auf einer Wort-zu-Wort-Basis, wobei rasch erkannt wurde, dass dies angesichts der unterschiedlichen grammatikalischen Natur der Sprache nicht gut funktionierte.

Beispiele von Bereichen der Computerlinguistik, die der Massenüberwachung zuspielen und wo Einsatzzwecke für die Massenüberwachung sichtbar sind, können auf Basis von Forschung, die in der EU (*INDECT*) oder USA (*NSA*) betrieben wurde, beispielhaft sein:

- Named Entity Recognition *NER*
- Text Mining *TM*
- Relation Extraction *RE*
- Text Summarization *TS*
- Topic Analysis *TA*
- Machine Translation *MT*

In einem Vortrag am 30. Chaos Communication Congress zeigt Joachim Scharloth exemplarisch auf, wie "politischer Extremismus" diverser in Deutschland populärer Blogs gemessen werden kann. Einerseits werden dafür Wortkollokationen eingesetzt, dann aber auch Verdachtsmomente in der Sprache auf Grund von Gradpartikeln ausgemacht, wie sie im Überwachungsmodell 2 dieser Arbeit ebenfalls eingesetzt werden. Durch seinen Vortrag postuliert er, dass Geheimdienste solche und ähnliche Verfahren sicherlich einsetzen, um ein Monitoring von "politischem Extremismus" zu betreiben. [40, 99, 100, 101, 107]

In den folgenden Abschnitten soll aufgezeigt werden, dass der Stand der Technik in der Computerlinguistik genutzt wird, um ihn für Zwecke der Massenüberwachung einzusetzen oder auf Basis bestehender Erkenntnisse neue CL-Modelle zu erstellen, die spezifisch dafür entwickelt werden.

2.2.2 EU: INDECT-Projekt

“Intelligent information system supporting observation, searching and detection for security of citizens in urban environment” *INDECT* bezeichnet ein Forschungsprojekt, das im Rahmen des 7. Forschungsrahmenprogramms der Europäischen Union *EU* von 2009 bis 2013 erforscht wurde. Für den angegebenen und übergeordneten Zweck, die Sicherheit aller Bürger in der *EU* zu erhöhen, wurden in neun Work-Packages[33] zahlreiche Methoden – auch der Bild-, Biometrie- und Videoanalyse – erforscht und evaluiert, die in eine umfassende und systemische Plattform der Massenüberwachung münden könnten: die *INDECT*-Plattform. Jedes Work-Package hat andere Schwerpunkte, wobei sich insbesondere das Work-Package 4 mit der Erforschung von Methoden aus der Computerlinguistik beschäftigt. [65]

Erforschte Methoden, die als genuin computerlinguistisch betrachtet werden können, sind in der Tabelle 2.1 zusammengefasst.

Gebiet	**Neutraler Zweck**	**INDECT-Zweck**	**INDECT-Referenz**
Named Entity Recognition *NER* / Relation Extraction *RE*	Erfassung bestimmter namentlicher Entitäten *NE* und Herstellen von Beziehungen zwischen diesen	Finden von Verdächtige (beispielsweise “Hooliganismus” oder “Terrorismus”) in Chats oder Foren	Klapaftis (47, 50), Klapaftis, Ioannis P. and Nagy, Zoltán and Johanning, Nils (52), Klapaftis (49)
Information Retrieval *IE* / Pattern Matching / Text Mining *TM*	Suchstrategien nach Relevanz (gegebenenfalls mit spezifischen sprachlichen Suchmustern)	Finden von verdächtigen Inhalten im Web (“Drogenhandel”, “Waffenhandel”, “terroristische” Aktivitäten und Rekrutierung)	Dorosz et al. (17), Michał und Lubaszewski (70), Dorosz (16), Jung Pandey und Dorosz (43), Pandey und Dorosz (91), Dorosz et al. (18), Korzycki und Lubaszewski (55), Pandey (88)
Machine Learning *ML*	Erkennung von Regelmässigkeiten in Daten; Treffen von Voraussagen; Kategorisierung	Erkennung von Verdächtige auf Grund ihres sprachlichen Ausdrucks (Profilierung nach Verhalten)	Klapaftis (48), Klapaftis und Pandey (51), Pandey (89)

Table 2.1: Übersicht INDECT-Forschung: Methoden der Computerlinguistik

Weiterhin sind auch Forschungsergebnisse vorhanden, wo computerlinguistisches Wissen im Zusammenhang mit Erkenntnissen anderer Disziplinen zusammenwirkt. Wenn es beispielsweise darum geht

- Personen auf Grund auditiver Merkmale nach Geschlecht, Alter oder nach ihrer emotionalen Verfassung hin zu erfassen und zu profilieren, indem beispielsweise festgestellt wird, dass ihr Stottern ein Ausdruck von Nervosität ist; [92]

[33] Das letzte Work-Package besteht aus Arbeiten der Evaluation vorangegangener Forschung.

- eine Ontologie aufzubauen, die es erlaubt, die Welt formal zu beschreiben, und damit regel- und datenbasiert Handlungsstrategien für Polizei, Bahnverkehr und andere Anspruchsgruppen zu ermöglichen, wenn es zu Krisensituationen kommt; [117]
- oder Systeme zu betreiben, wo User-Generated-Content *UGC* in strukturierter und unstrukturierter Form aufgenommen werden kann, und die Verwaltung und der Abruf dieses (natürlichsprachlichen) Wissens zu organisieren sind. [61, 90]

Im Zusammenhang mit dieser Arbeit sind auch Forschungsergebnisse interessant, welche konkret mit der Überwachung von Datenströmen zu tun haben und was daraus alles gewonnen werden kann.

So wird in einem Paper mittels *Xplico*[34] und Techniken der Deep-Packet-Inspection, verdächtiges Verhalten beim Websurfen entdeckt: konkret werden Bilder im Internetverkehr mitgeschnitten und mit solchen einer zentralen Datenbank, die wiederum als verdächtig markierte Bilder enthält, verglichen. Dies resultiert in einem Alert, der schliesslich von einem Agenten manuell begutachtet werden kann. Als Vergleichsbasis kommen Hashsummen zum Einsatz.[35] [27]

Es werden zudem weitere Gebiete erforscht, wo Computerlinguistik zumindest am Rande eine Rolle spielen kann:

- Die Analyse der Zugehörigkeit von Personen zu Gruppen in Blogs.[36] [29]
- Die agenten-basierte Modellierung von Gesellschaft wird betrieben, was makroskopisch betrachtet, viel mit Modellen aus der Informatik, Soziologie oder Mathematik zu tun hat, auf individueller Ebene (von Sprechakten) allerdings computerlinguistisches Wissen erfordern kann. [56]
- Die Handhabe der verschiedenen Dateiformaten, die im INDECT-Projekt zusammenlaufen, wird erforscht. [128]

2.2.3 USA: HLT-Programm der NSA

Im Rahmen des Human Language Technology *HLT*-Programmes der *NSA* wird Sprachtechnologie entworfen, die direkt zum Zweck der Massenüberwachung eingesetzt wird. Zur Verfügung stehen jährlich dutzende Millionen USD, welche dieser *NSA*-eigenen Forschung zufallen. [78]

Im Mai 2015 wurden im *The Intercept* NSA-Dokumente veröffentlicht [24], welche aufzeigen,

[34] Ein Werkzeug zur Überwachung von Datenströmen.

[35] Das ermöglicht einerseits zwar einen effizienten Vergleich, andererseits funktioniert das bei (leicht) veränderten Bildern nicht zuverlässig, weil sich die Hashsummen grundlegend unterscheiden.

[36] Hier können Metadaten der Kommunikation eine wichtige Rolle spielen, wie sie im Fokus der Arbeit von Daniel Mossbrucker stehen. [72]

dass Speech-To-Text *STT* ein Gebiet ist, das die NSA bearbeitet; über die transkribierten Daten soll sodann extensive Suche auf Basis von Keywords[37] erfolgen, wenn auch die dafür erforderliche "perfekte Transkription" aus den mitgeschnittenen Tonsignalen als Basis dafür, auch bei den Gehemdiensten nicht gelöst ist:

> *Though perfect transcription of natural conversation apparently remains the Intelligence Community's "holy grail" , the Snowden documents describe extensive use of keyword searching as well as computer programs designed to analyze and "extract" the content of voice conversations, and even use sophisticated algorithms to flag conversations of interest.*

Für diesen Bereich der Voice-Kommunikation besteht konkret ein Programm namens *RHINEHART*, das es erlaubt, Transkriptionen gesprochener Sprache durchsuchbar zu machen. [119]

Besonders interessant erscheint das mitveröffentlichte "Classification Guide"-Dokument, das zur Einstufung der HLT-Modelle der *NSA* dient. Darin ist sichtbar, dass der Umstand, dass die *NSA* sprachtechnologische Modelle entwickelt zwar nicht geheim ist, die konkreten Modelle allerdings schon: damit wird verhindert, dass Personen wissen können, nach welchen sprachlichen Auffälligkeiten oder Mustern die *NSA* genau sucht. Im fremdsprachlichen Bereich scheint ein Schwerpunkt auf die Analyse von arabisch-, doch auch spanischsprachigen Inhalten zu liegen. Die spanischsprachigen Modelle sollen sich in den letzten Jahren stark verbessert haben. [79, 84]

In Folien [77] des "Center for Content Extraction" wird ein *STAIRS* genanntes Programm erwähnt, das in folgenden Schritten funktioniert:

1. Mittels inhaltlicher Selektoren wird verdächtiger Text ausgewählt.
2. Es erfolgt eine Übersetzung auf Basis eines "Glossars".
3. In Wikis festgehaltenes Wissen wird beigezogen, um das gefundene Material (semantisch) einzuordnen.

2.2.4 USA: NSA-Patente mit Computerlinguistik-Bezug

Im Rahmen von Recherchen von *NSA*-Patenten, welche computerlinguistisch relevant und für Massenüberwachung geeignet sind, wurden 13 Patente, die zwischen 1995 und 2013 akzeptiert wurden, gefunden.[38]

Sie betreffen zumindest die folgenden Gebiete:

[37] Soft-Selektoren

[38] Ein Anspruch auf Vollständigkeit besteht nicht.

- Methoden des Information Retrieval *IE* oder der Datenbanksuche[39]
- Generierung von Selektoren als n-gramm-Muster, "Keywords", "index terms" (Wortkombinationen möglich) und ähnlichen Bezeichnungen (aus kontinuierlichen Textströmen[40]) [41]
- Methoden zur automatischen Textzusammenfassung[42], teilweise unter Einsatz von Topic Analysis *TA*[43]
- Methoden der Topic Analysis *TA* [44]
- Methoden der Optical Character Recognition *OCR*[45]
- Methoden der Machine Translation *MT*[46]
- Methoden, sprachliche (Voice-)Kommunikation zu visualisieren[47]

Zahlreiche der Technologien, welche die *NSA* patentiert hat, stellt sie im Rahmen eines Technology Transfer Programs *TPP* interessierten Kreisen für die Anwendung zur Verfügung, sofern diese als Lizenznehmer akzeptiert werden und sich verpflichten, Geheimhaltung über die Technologie und dem Einsatz des jeweiligen Produkts zu wahren. [83]

2.3 Im Fokus: Sprachtechnologische Produkte zur Massenüberwachung

In diesem Teil der Arbeit wird auf zwei sprachtechnologische Produkte fokussiert, die – dokumentiert – dafür existieren und angepriesen werden, konkret auf natürlichsprachlicher Basis Massenüberwachung durchzuführen.[48]

2.3.1 Snowden-Enthüllungen: XKeyscore

XKeyscore (im Folgenden *XKS*) ist eine mächtige Suchmaschine, die weit über die Fähigkeiten von bekannten Websuchmaschinen[49] hinaus, dafür genutzt werden kann, insbesondere

[39]Vgl. im Anhang die Abstracts B.1, B.3, B.9, B.10.
[40]Engl. "continuous text streams"
[41]Vgl. im Anhang die Abstracts B.2, B.3, B.4 und B.5.
[42]Engl. Text Summarization *TS*
[43]Vgl. im Anhang die Abstracts B.6 und B.7.
[44]Vgl. im Anhang den Abstract B.11.
[45]Vgl. im Anhang den Abstract B.8.
[46]Vgl. im Anhang den Abstract B.12.
[47]Vgl. im Anhang den Abstract B.13.
[48]Auch wenn sie über weitere Analysefähigkeiten verfügen, die nicht aus der Computerlinguistik kommen.
[49]Wie sie für die Evaluation im Kapitel 4 eingesetzt werden.

private Kommunikationsströme und -datenbestände, die weltweit abgefangen und zugeführt wurden, mit (komplexen) Selektoren und sogenannten "fingerprints" zu durchsuchen. Letztere sind insbesondere Filter für bestimmte Dienste. Die durchsuchbaren "Korpora" werden aus restlos allen Quellen, wie sie unter 2.1.2 beschrieben werden, aufgebaut.

Zugriff auf die weltweit verteilt arbeitenden *XKS*-Systeme haben Geheimdienste aus dem *FVEY*-Verbund, gemäss Medienberichten und Ergebnissen aus dem *NSAUA* aber auch andere (bis anhin enge) Partnerdienste wie der deutsche Auslandsgeheimdienst *BND* – seit 2007. Zudem gibt es seit August 2015 Belege dafür, dass *XKS* auch beim deutschen Inlandsgeheimdienst Bundesamt für Verfassungsschutz *BfV* mit Codename *Poseidon* im Einsatz ist. [3]

XKS wird in der *FAZ* [57] wie folgt charakterisiert:

> *Auch in der vergangenen Woche kamen wieder neue Details ans Licht darüber, wie mit dem schon bekannten NSA-Programm XKeyscore private und geschäftliche Daten zur Analyse weiterverarbeitet werden. Denn nach wie vor durchforsten und sammeln die mehr als 700 über die Welt verteilten XKeyscore-Server unaufhörlich neues Material: So landen jeden Tag Hunderttausende Sprach- und Textnachrichten, alle Arten von E-Mail-Anhängen, Bilder aus Webcams, Skype-Anrufe oder Passwörter in den NSA-Datenbanken und werden in den wachsenden Datenhalden prozessiert. Die Grössenordnung der Sammlung musste nochmals noch oben korrigiert werden. Hinzu kommt die Erkenntnis, dass die Hacking-Operationen in industriellem Massstab konzeptioniert, durchgeführt und ausgewertet werden.*

Am 1. Juli 2015 wurden im *The Intercept* [66] zahlreiche Folien mit Bezug zu *XKS* veröffentlicht und die Fähigkeiten von *XKS* noch einmal rekapituliert. In einer Auswahl sind folgende Einsatzzwecke von *XKS* dokumentiert:

- Dank dem britischen *Tempora*-Programm, das Daten aus Unterseekabeln vor der britischen Küste abfängt und speichert, kann *XKS* dazu genutzt werden, mindestens über drei Tage in Inhalts- und 30 Tage in Metadaten zu suchen, die in Europa ein- und ausgehen.
- Besucher "extremistischer" Webseiten können erkannt und mit all ihren Telekommunikationsmerkmalen überwacht werden.
- Techniken zur Extraktion von Telefon-Nummern oder anderen Telekommunikationsmerkmalen, die der Form nach Metadaten sind, und als formale Selektoren (später) eingesetzt werden können, werden geschildert.
- Ein Projekt, um Akteure in Webforen[50] zu beobachten und ihre Fähigkeiten zu erfassen, wird beispielhaft erwähnt.

[50] Im Beispiel: Hacker-Webforen

- Für sogenannte Tailored Access Operations *TAO* können “fingerprints” eingesetzt werden, die es erlauben, vulnerable Computersysteme im Internet zu finden, um sie als zusätzlich durchsuchbare Datenquellen nutzbar zu machen. Dafür werden diese im Rahmen von *TAO*-Operationen angegriffen und “gehackt”.
- Der HTTP-Traffic in Datenströmen (oder analog auch der anderer Protokolle) kann minutiös analysiert werden: so kann beispielsewise festgestellt werden, welche Webbrowserkennungen genau im Einsatz sind oder wie ein Webseitenbesucher über Links im Web navigiert. Das erlaubt einerseits die weitergehende Profilierung von Webseitenbesuchern, allerdings auch Angriffe auf den beobachteten Akteur auszuführen.
- Die diversen Typen von Selektoren und wie sie eingesetzt werden können, werden illustriert. Spezifisch wird auch kontextsensitive Suche gezeigt, die ohne Strong-Selektoren auskommt, wo inhaltliche Suchbegriffe eingesetzt werden, wie sie im praktischen Teil dieser Arbeit im Fokus stehen.

Nutzt man die *XKS*-Syntax nicht korrekt, können schnell falsche Personen in den Fokus geraten; zudem können ganze IP-Ranges abgescannt werden, wie im “Unoffical XKS User Guide” geschildert wird. [82, S.24]

2.3.2 WikiLeaks-Publikation: Search & Detect

Im Rahmen der *The Spy Files*-Veröffentlichungen[51] von *WikiLeaks* kam es zur Publikation zahlreicher Materialien, welche vor allem aufzeigen, dass auch in der Privatwirtschaft eine Fülle an fertigen Produkten existiert, um im grossen Massstab eine Massenüberwachung von Kommunikation zu betreiben.

Besondere Aufmerksamkeit und hier im Fokus der Darstellungen verdient das sprachtechnologische Produkt *Scan & Target* der gleichnamigen Firma, das in Frankreich hergestellt wurde.[52]

Das Produkt verspricht sowohl Social-Media, als auch Foren, Blogs, E-Mails und IM-Dienste in Echtzeit und multilingual[53] durchsuchen zu können sowie auf Basis von erkannten Mustern Alarmierungsfunktionen anzubieten. Unter den Kunden figurieren privatwirtschaftliche Akteure wie L’Oréal oder IKEA – ohne allerdings konkret zu werden, wie diese Unternehmen diese Sprachtechnologie genau nutzen. Interessanter ist sodann das Angebot auch für den Geheimdienstbereich der “intelligence” sowie Strafverfolgungs-

[51] WikiLeaks (2011). *The Spy Files*.
URL https://wikileaks.org/the-spyfiles.html. Abruf: 13. November 2015.

[52] Scan & Target (2011). *Extracting intelligence from multilingual SMS, IM, e-mails. . .*
Online: https://wikileaks.org/spyfiles/files/0/71_201110-ISS-IAD-T5-SCAN_AND_TARGET.pdf
(Abruf: 13. November 2015)

[53] Unterstützt werden Englisch, Französisch, Spanisch und diverse arabische Dialekte inklusive Transliteration.

behörden offen zu sein und hier OSINT- und COMINT-Fähigkeiten anzubieten: mitsamt einer API, die es erlauben soll, die Software für Big-Data-Anwendungen einzusetzen. Ein Beispiel im Bereich OSINT wird mit der Social-Media-Plattform Twitter angeführt: das Produkt sei unter Einsatz entsprechender IBM-Hardware fähig, den gesamten Traffic des Kurznachrichtendienstes von (damals: 2011) rund 10TB je Tag zu verarbeiten und automatisch massen zu überwachen.[54]

In computerlinguistisch interessantester Hinsicht wird darauf hingewiesen, dass das Produkt in den unterstützten Sprachen folgende Fähigkeiten (und mehr) aufweist:

- Es wird zwischen Gross- und Kleinschreibung differenziert.
- Bewusster Einsatz von oder durch Flüchtigkeitsfehler entstandene Buchstabenrepetitionen, die den semantischen Zusammenhang in einer Textkollektion mildern könnten, werden beachtet: `vvvvvviagra` = `viagra`.
- Ortografische Schreibvariationen werden beachtet: `viagra` = `vi@gra` = `vlarga`.
- Geringe linguistische Distanzen von vorgefundenen Oberflächenformen zum Lemma können korrigiert werden, wenn einzelne Buchstaben fehlen, eine (falsche) Worttrennung oder Substitution vorliegt: `viagra` = `v iagra` = `v|agra`.
- Wortalterationen unter Einsatz nicht-alphanummerischer Zeichen können erkannt werden: `viagra` = `v.i.a.g.r.a` = `v-iagra`.
- Auf die diversen Textsorten wird eingegangen: dem (informelleren) Schreibstil in SMS- oder IM-Konversationen und den kleineren Dokumentengrössen wird Rechnung getragen.
- NER wird betrieben zur Erkennung zum Beispiel – gemäss Folien – von Namen, Orten oder Nationalitäten.
- Konversationsfäden können aufgeschlüsselt und das soziale Netzwerke der Kommunikationsteilnehmer erfasst werden.

Ohne, dass in den Folien selber auf die konkreten computerlinguistisch zugrundeliegenden Methoden eingegangen wird, führen die Autoren an, dass die überwachte Kommunikation nicht bloss auf Einzelwortbasis analysiert wird, sondern Kontext und Bedeutung miteinfliessen; so sei das Produkt fähig, die analysierten Textströme nach Topics[55] zu "filtrieren" und ferner Sentimentanalyse zu betreiben oder Fragen erkennen zu können.

In den Folien sind ferner zwei konkrete Fallbeispiele der Inhaltsanalysefähigkeiten gegeben:

- Der Umgang mit verschiedenen Arabischdialekten und -schreibweisen ist handhabbar:

[54] Zu betonen ist, dass der Begriff der "mass interception" (zu Deutsch: Massenüberwachung) in den Folien unverblümt selber geführt wird.

[55] Es werden die Bereiche Terrorismus, Drogen oder Gewalt genannt.

das Produkt ist fähig, das arabische Chat-Alphabet mit den phonetisch äquivalenten Zeichen zum arabischen Alphabet zu erfassen, die in der Textkommunikation über das Internet Anwendung findet.

- Das illegale Anbieten von kinderpornografischen Darstellungen wird auf Grund einer offenbar szenetypischen Abkürzung und Alterserwähnung (eines Kindes) im Zusammenhang mit der Verlinkung eines Medienerzeugnisses beispielhaft erkannt.

Es wird behauptet, die False-Positive-Rate liege im Bereich $[0.001, 0.05]$: dies kann allerdings mangels einer wissenschaftlichen Evaluation im Rahmen der verfügbaren Dokumente nicht unabhängig überprüft werden, schon gar nicht, wenn nicht deutlich wird, in welchen Kontexten dies stimmen soll.

2.4 Befunde: Logiken der Massenüberwachung auf Basis von Selektoren

2.4.1 Was sind Selektoren?

Gemäss *BBC*-Glossar [45] zu den NSA-GCHQ-Enthüllungen können Selektoren wie folgt definiert werden:

> *The identifiers used by the security services when carrying out their searches - for example, a phone number or an internet protocol (IP) address. These can be divided into two broad types – strong and soft.*
>
> *A strong selector is an identifier associated with a specific individual – e.g. a search for content associated with an email address. In theory, the NSA cannot intentionally use a strong selector linked to a US citizen, any other US person or anyone located within the United States.*
>
> *A soft selector is less limited in its scope and is typically based on the content of message. This can be a word or phrase, such as "big explosion", or the language the message is written in – for example, French.*

Das heisst, dass zwischen "Strong"- und "Soft"-Selektoren unterschieden wird, entsprechend ob nach Meta- oder Inhaltsdaten der Kommunikation aussortiert werden soll.

Grundsätzlich wird diese Unterscheidung von allen Staaten, die Massenüberwachung betreiben und die im Folgenden betrachtet werden, gemacht, wenn auch die Bezeichnungen unterschiedlich sein können.

2.4.2 FVEY: Echelon

> *Although it is impossible for analysts to listen to all but a small fraction of the billions of telephone calls, and other signals which might contain "significant" information, a network of monitoring stations in Britain and elsewhere is able to tap all international and some domestic communications circuits, and sift out messages which sound interesting. Computers automatically analyse every telex message or data signal, and can also identify calls to, say, a target telephone number in London, no matter from which country they originate.*
>
> – Duncan Campbell am 12. August 1988 [14]

1988 enthüllte der investigative Journalist Duncan Campbell das Echelon-System, ein globalisiertes Überwachungssystem, das von den FVEY-Staaten begründet wurde. Bereits für damalige Verhältnisse mag erstaunen, dass das System fähig war, etwa Milliarden von Telefonanrufen nach "signifikanten" Informationen automatisch auszusondern. [96, S.107] [14]

Offiziell bestätigt und als amtlich existent [105] gilt *Echelon* seit 2001, nachdem Gerhard Schmid, Berichterstatter des Europäischen Parlaments, in einem umfassenden Bericht Wesen und Fähigkeiten von *Echelon* darlegte. Der Bericht wurde mit qualifizierter Mehrheit angenommen.

Schmid qualifiziert Spionage als nicht Geringeres als den *"organisierten Diebstahl von Informationen"*, wobei diese bei *Echelon* insbesondere durch einen Abhörverbund der UKUSA-Staaten weltweit organisiert ist. Unter anderem betont er die Rolle des Vereinigten Königreichs, welche die Fähigkeit hat und nutzt, die Unterseekabel vor der eigenen Küste anzuzapfen. [105, S.28,32,34] Dies hat sich mit dem Snowden-Enthüllungen konkret bestätigt, indem das *Tempora*-Programm enthüllt wurde.

Zur Art der Suche in den Datenströmen wird im Bericht festgehalten, dass diese *"[...] auf Basis von Suchbegriffen, die thematisch gruppiert sind [...]"* erfolge – ohne allerdings auf die Genese der Suchbegriffe einzugehen. Am Beispiel des Bundesnachrichtendienstes *BND* zeigt Schmid auf, dass von den (damals: 2001) rund 10 Millionen internationalen Kommunikationsverbindungen, die je Tag über Deutschland erfolgen, 800'000 über Satellit abgewickelt würden: davon würden 75'000 der Rasterung durch eine Suchmaschine unterworfen. Als Grund, weshalb nur knapp 10% des satellitengestützten Verkehrs durchsucht wurde, gelten technische Beschränkungen. Es wird ferner darauf hingewiesen, dass wenn zusätzlich Kabelverkehr beigezogen wird, zwar die statistische Trefferwahrscheinlichkeit steigt, weil das durchsuchbare Textkorpus grösser ist, das aber nicht unbedingt heisst, dass deswegen auch mehr True-Positives resultieren. Für *Echelon* vermag der Berichterstatter keine Angaben über die Zahl der verfügbaren Begriffe zu machen. Dafür liegen solche für den *BND* vor, der – wie wir heute wissen – eng mit dem *FVEY*-Verbund und insbesondere der *NSA* zusammenarbeitet, so dass es sich möglicherweise um ähnliche, wenn nicht gar

dieselben inhaltlichen Suchbegriffe handelt. Es wird festgestellt [105, S.37,38], dass es je nach Themenbereich eine verschieden grosse Zahl an inhaltlichen Selektoren gibt:

- 2'000 für den Bereich Proliferation
- 1'000 für den Rüstungshandel
- 500 für den Terrorismus
- 400 für den Drogenhandel

Für die letzten zwei Bereiche hält Schmid [105, S.38ff.] fest:

> *Bei Terrorismus und Drogenhandel hat sich das Verfahren allerdings als nicht sehr erfolgreich erwiesen.*

Für andere Überwachungsbereiche liegen weder für den *BND* noch *Echelon* genaue Informationen vor, allerdings wird darauf hingewiesen, dass einerseits jeder Mitgliedstaat des *FVEY*-Verbundes über seine eigenen "Dictionary"-Computer verfüge, die es mit beliebigen Suchbegriffen ausstatte könne, um Texte der interessierenden Suchkategorie zu finden und andererseits seit Anbeginn des Verbundes[56] das Interesse auch den Bereichen Wissenschaft und Wirtschaft und nicht bloss Politik und Sicherheit gilt. Diese Aussage werden mit Belegen durch *Echelon* durchgeführter Wirtschaftsspionage untermauert. [105, S.70ff.,108ff.]

Die Schweizer Geheimdienstkontrolle hat sich ebenfalls *Echelon* angenommen. Sie gelangt 2003 zu ähnlichen Erkenntnissen: dass System umfasse nicht nur die Überwachung von Funkverkehr, sondern auch von Kabelnetzen; zudem sei das globalisierte COMINT-System ursprünglich zwar militärisch zu verorten, rasch aber wurde es für Zwecke der *"Wirtschafts- und Konkurrenzspionage"* eingesetzt. [30, S.1507ff.]

Es wird im Bericht auch auf Gefahren für den einzelnen Bürger hingewiesen: so soll durch *Echelon* eine Frau als potenzielle "Terroristin" markiert worden sein, weil sie in einem Telefongespräch mit einem Freund einen zweitdeutigen Begriff genutzt habe. [105, S.75]

Selbst entgegen den Interessen der Regierungen des *FVEY*-Verbundes selber soll *Echelon* bereits missbraucht worden sein, wie folgendes Zitat [105, S.75] zu bedenken gibt:

> *[D]er GCHQ [habe] den kanadischen CSE gebeten, für ihn zwei englische Minister auszuspionieren, als Premierministerin Thatcher wissen wollte, ob diese sich auf ihrer Seite befinden.*

So empfiehlt [105, S.133] das *Europäische Parlament* mit der Verabschiedung des Berichts abschliessend auf Technologien der Mailverschlüsselung zu setzen, und darauf bedacht zu

[56]Das UKUSA-Agreement, das den den *FVEY*-Verbund begründet, wurde nach Ende des Zweiten Weltkriegs eingerichtet.

sein, interne Netze möglichst vom Internet zu trennen: Empfehlungen, die – wie wir aus den Snowden-Enthüllungen wissen – heute mehr denn je unterstreicht werden können.

Technisch konkreter hinsichtlich der Fähigkeiten von *Echelon* wird ein technischer Bericht der Europäischen Union, der bereits 1999 erschienen [1, S.53ff.] ist. Die Suche nach "keywords" wird mit der einer Websuche verglichen, allerdings mit unvergleichbar grösseres Korpus:

> *Whenever machine readable communications are available, keyword recognition is fundamental to Dictionary computers, and to the ECHELON system. The Dictionary function is straightforward. Its basic mode of operation is akin to web search engines. The differences are of substance and of scale. Dictionaries implement the tasking of their host station against the entire mass of collected communications, and automate the distribution of selected raw product.*

Weiterhin gibt es nebst der "keyword"-basierten Suche auch Verfahren, die n-gramm-basiert nach Themen suchen. Das System wird wie folgt beschrieben:

> *To use N-gram analysis, the operator ignores keywords and defines the enquiry by providing the system with selected written documents concerning the topic of interest. The system determines what the topic is from the seed group of documents, and then calculates the probability that other documents cover the same topic.*

In einem Test hat sich dieses Verfahren als besser erwiesen; es sollte dabei nach Evidenz von *Airbus*-Tochtergesellschaften in Textkollektionen gesucht werden:

> *In a standard US test used to evaluate topic analysis systems, one task the analysis program is given is to find information about "Airbus subsidies". The traditional approach involves supplying the computer with the key terms, other relevant data, and synonyms. In this example, the designations A-300 or A-320 might be synonymous with "Airbus". The disadvantage of this approach is that it may find irrelevant intelligence (for example, reports about export subsidies to goods flown on an Airbus) and miss relevant material (for example a financial analysis of a company in the consortium which does not mention the Airbus product by name). Topic analysis overcomes this and is better matched to human intelligence.*

Das Verfahren entspricht zwar nicht der Topic Analysis, wie es von Blei 2003 [8] später beschrieben wird, allerdings wird damit schon früh der Logik gefolgt, auf Basis eines Verdachtskorpus nach (weiteren) ähnlichen Texten zu suchen – wie dies auch im Rahmen des praktischen Teils dieser Arbeit umgesetzt wird.

Im Zweifelsfall (bei verschlüsselter Kommunikation) oder im Falle von Voice-Kommunikation würde "traffic analysis" eingesetzt, womit – in womöglich wichtigster Hinsicht – Kontaktnetzwerke von Personen ausgespäht werden können, was einer Massenüberwachung mit Fokus auf Metadaten entspricht:

Traffic analysis can be used where message content is not available, for example when encryption is used. By analysing calling patterns, networks of personal associations may be analysed and studied. This is a principal method of examining voice communications.

2.4.3 Deutschland: NSAUA zur Massenüberwachung des BND

Ausspähen unter Freunden – das geht gar nicht.

– Bundeskanzlerin Angela Merkel am 24. Oktober 2013 [86]

Auf Grund der Snowden-Enthüllungen wird die Beteiligung Deutschlands an der weltweiten Massenüberwachung analysiert: sie ist durch starke mediale Begleitung geprägt, die als Belege für einige Sachverhalte dienen können. Da der noch immer arbeitende parlamentarischen Untersuchungsausschusses [13] einige Details zur Funktionsweise der Massenüberwachung auf Basis von Selektoren zu Tage fördert, werden Teile ihrer Ergebnisse – sofern sie technische Aspekte betreffen – im folgenden Abschnitt 2.4.3.1 beleuchtet.

2.4.3.1 Zahl von Selektoren

Die Anzahl Selektoren, die beim deutschen Auslandsgeheimdienst *BND* im Einsatz sind oder waren, reichen je nach Zeitspanne der Betrachtung und laufender Überwachungsprogramme der Massenüberwachung von 800'000 bis zu 14 Millionen, wobei die meisten dieser Selektoren nicht vom *BND* selber stammen, sondern – im Zuge einer Joint SIGINT Activity *JSA* – von der *NSA*, teilweise ohne jede Kontrolle des *BND* über deren Natur oder Zweck, stammen. [4, 69]

So schreibt [93] Roland Peters unter dem Titel "Blinde Spionage des BND" für *N-TV*:

Es gibt [...] Informationen darüber, wie sorglos der deutsche Geheimdienst wohl einfach alles durchwinkte, was die Amerikaner filtern wollten. Den Zahlen des NSA-Untersuchungsausschusses im Bundestag zufolge liefen allein im August 2013 zwischen 8 und 9 Millionen Selektoren ein. Darin suchte der BND nach Kollisionspunkten mit deutschen Interessen. Er fand 25.000 und löschte sie.

2.4.3.2 Graulich-Bericht

Im Graulich-Bericht [36] zu Handen des *NSAUA* liefert der Berichterstatter Kurt Graulich, der die Aufgabe erhalten hat, 40'000 NSA-Selektoren zu überprüfen, folgende Erkenntnisse:

- Die Natur der Massenüberwachung wird bestätigt, wenn der Berichterstatter es auch vorzieht, bloss von der Überwachung einer *"Vielzahl"* von Telekommunikationen zu

schreiben.

- Für einen formalen Suchbegriff (oder: Strong-Selektor) sind bis zu 20 Varianten (*“Permutationen”*) möglich.
- Inhaltliche Suchbegriffe (oder: Soft-Selektoren) würden auch eingesetzt, seien allerdings seltener, weil sie zu einer hohen Zahl irrelevanter Treffer führen würden.
- Auf eine Suchabfrage erfolgen bis zu drei weitere (interne) Suchen, die als “Filterstufen” gelten und die Trefferstatistik negativ beeinflussen; diese sind:
  1. Liste mit eindeutig deutschen Grundrechtsträgern, die nicht überwacht werden dürfen.
  2. Liste mit deutschen Grundrechtsträgern (Personen, Firmen oder andere Entitäten), die im Ausland verweilen, und auch nicht überwacht werden sollen.
  3. Liste mit Selektoren, die darauf abzielen, “deutsche Interessen” zu wahren.
- Bei den formalen Selektoren, die beschrieben werden, sind eine Reihe von Selektortypen bekannt, wie Telefon-Nummern oder IP-Adressen; andererseits besteht auch eine Kategorie “OTHER”, für alle Telekommunikationsmerkmale, wo keine einfache Zuordnung möglich ist.
- Graulich konstatiert, dass die Interpretation der Selektoren schwierig ist: eine *“subjektive Komponente”* sei gegeben.
- Was die von *WikiLeaks* publizierten Selektoren [126] gegen die eigene Regierung betrifft, so wisse der *BND* davon nichts.

Der Bericht wird von netzpolitischen Akteuren allerdings umgehend auch kritisiert: so habe Graulich *SIP*-Kennungen für Internettelefonie mit E-Mail-Adressen verwechselt oder ein falsches Verständnis darüber, was eine *IMEI* ist oder einen allgemein viel zu verkürzten Begriff vom Wesen von Selektoren. Der Berichterstatter wird zudem dafür kritisiert, sich vom *BND* bei der Abfassung des Berichts unterstützt haben zu lassen, was die Unabhängigkeit des Berichts in Frage stelle. [7, 58]

So schreibt [58] Constanze Kurz vom *CCC* in der *FAZ*:

> *Nach der Lektüre von Graulichs Bericht drängt sich eine Erkenntnis förmlich auf: Das Problem sind nicht allein die Fehler, die Rechtsbrüche und das Durchwinken bei der Weitergabe von Informationen an die NSA. Das Problem ist die “strategische Fernmeldeüberwachung” an sich. So zu tun, als könnte man das massenhafte Durchforsten von Kommunikation durch die Anwendung von mehr oder weniger Selektoren irgendwie beherrschen und “grundrechtskonform” durchführen, ist als bloße Augenwischerei enttarnt.*

Was das Zustandekommen der Selektoren betrifft, herrscht weiterhin Unklarheit, wobei eine interessante Analyse in der Sache aus Österreich kommt: der dortige Parlamentsabgeordnete Peter Pilz verweist in einem Facebook-Post auf eine durch WikiLeaks veröffentlichte US-Depesche vom 31. Juli 2009[57] in der die Botschaften aufgerufen werden, HUMINT-Informationen von UN-Funktionären zu sammeln.

In dem Zusammenhang beschreibt er das Zustandekommen von US-Selektoren inmitten Europas:

> *Die ausspionierten HUMINT-Daten sind dann die Selektoren, mit denen die NSA ihre in Frankfurt und Wien abgeschöpften Massendaten durchsucht. Mit diesen Selektoren werden dann die österreichischen Daten, die durch die SCS-Antennen auf US-Botschaft und US-Mission, am Telekom-Knoten Frankfurt, über direktes Hacken einzelner Rechner und über Satelliten abgeschöpft wurden, durchsucht.*
>
> *So funktioniert der amerikanische Spitzelstaat mitten in Wien.*

2.4.3.3 Spionage von Freunden und "hochkritische" Selektoren

In jüngeren Erkenntnissen aus dem *NSAUA* wurde bekannt, dass der *BND* einen eigenen (deutschen) Diplomaten ausspioniert hat, daneben aber auch einige Minister befreundeter Staaten sowie deren Botschaften. Was die Schweiz betrifft, so soll unter anderem das Internationale Komitee vom Roten Kreuz *IKRK* in Genf Ziel von *BND*-Spionage gewesen sein. [37, 87]

Abgeordnete des *NSAUA* haben seit Ende November 2015 Zugang zu einer Selektorenliste im Bundeskanzleramt: erste Reaktionen deuten darauf hin, dass es darunter *"hochkritische"* Selektoren hat. Andere seien ohne Erklärungen nicht verständlich. In weiteren Fällen sei der Zusammenhang zu den bezeichneten Überwachungsgebieten (wie beispielsweise Proliferation oder Waffenhandel) nicht offensichtlich. [2]

2.4.4 Schweiz: Onyx oder Massenüberwachung von Militär und NDB

> *Der Einsatz des Onyx war gemäss spirituellen Kreisen – wie etwa dem Militärdepartement – absolut notwendig, um negative Einflüsse von bösen Menschen oder schwarzer Magie abzuwehren, und der Achat komplettierte diese Mission auf gar wunderbare Art und Weise. Kurzum: Der Achat war ein Glanzstück der nationalen Sicherheitspolitik, ein unverzichtbarer Bestandteil von Ueli Maurers Kronjuwelen, ein absolutes Muss für den Nachrichtendienst, wollte dieser kün-*

[57]Vgl. Publikation in The Guardian. [124]

> *ftig mit den Spitzenplayern auf dem Markt wie etwa NSA, KGB oder Mossad mithalten. Da durften auch die Anschaffungskosten von rund neunzig Millionen Franken kein Hindernis darstellen, Finanzengpässe hin oder her.*
>
> – Susi Stühlinger am 9. April 2015 [115]

1999 recherchierte die *SonntagsZeitung*, dass die Schweiz ein Funküberwachungssystem *Onyx* aufbaut, im Sinne vom *“[grossen] Lauschangriff im All”*. Zusätzliche Quellen deuten darauf hin, dass die angeschaffte Technik und ihre Ausrichtung, dafür eingesetzt würde, unter anderem deutsche, französische oder italienische Telekommunikationssatelliten abzuhören. [114] Bewilligt wurde das Initialbudget über CHF 45 Millionen von *Onyx* an mehreren Schweizer Standorten vom Parlament als “Mehrzweckgebäude”. Dem zuvor ging ein geheimer, unprotokollierter Beschluss des Schweizer Bundesrates, ein Abhörsystem nach dem Vorbild von *Echelon* zu errichten. [22, 98]

Tatsächlich hat *Onyx* erst 2012 nach rund 10-jährigem Betrieb gesetzliche Grundlagen überhaupt erhalten. [34, S.6586][35, S.3574]. Seither ist die Funkaufklärung im Bundesgesetz über die Zuständigkeiten im Bereich des zivilen Nachrichtendienstes *ZNDG* existent.

Wie die Weltwoche 2005 beschreibt [22], ähnelt die Funktionsweise dem *FVEY*-Vorbild durchaus:

> *Die Resultate des Systems dürfen nicht unterschätzt und als Spielerei abgetan werden. Auf den Onyx-Grossrechnern laufen Programme, welche alle abgesaugten Rohinformationen mit Hilfe von Künstlicher Intelligenz (KI), optischer Texterkennung (OCR), Sprach- und Stimmprüfung sowie von Schlüsselwort- und Themenanalysen filtern und sortieren. Werden vier bis fünf dieser “hitwords” oder “keywords” kombiniert, lässt sich die riesige Datenflut entscheidend kanalisieren. [...] Nach Meinung des deutschen Geheimdienstexperten Erich Schmidt-Eenboom kann mit dieser Methode der undurchschaubar scheinende Informationswirrwarr auf eine überschaubare Menge reduziert werden: “Werden die Suchbegriffe gezielt verbunden, reduziert sich die gigantische Informationsmasse rasch auf ein politisch und polizeilich verwertbares Mass.” Von der hochgeheimen Liste dieser Schlüsselwörter, erstellt von den Geheimdiensten, offiziell abgesegnet vom Bundesrat, weiss man nur, dass sie “laufend aktualisiert” wird. Die Liste der “hitwords” im Bereich des Waffenhandels soll mehr als zehn Seiten mit je 25 Begriffen umfassen.*

2.4.4.1 Untersuchungen der GPDel zur Funkaufklärung

Schon vor rund 12 Jahren untersuchte[58] die Geschäftsprüfungsdelegation der Eidgenössischen Räte[59], die in der Schweiz für die Geheimdienstkontrolle zuständig ist, das Funküberwachungssystem "Onyx", das *"[...] eine Massenüberwachung von Kommunikationen [ermöglicht] [...]"*, die dem Bereich COMINT zuzuordnen ist, die diskursive Kommunikationen betrifft: der Fokus liegt auf die *"[...] Abhörung, Auswertung und Übermittlung von Funkausstrahlungen, die in Graphiken oder in die menschliche Sprache übersetzt werden können [...]."* [30, S.1500,1505ff.]

Im Kern stellt die *GPDel*, was die Funktionsweise der Massenüberwachung betrifft, Folgendes fest:

> *Die Leistungsvereinbarungen beinhalten sämtliche zur Ausführung und Kontrolle der Aufträge erforderlichen Elemente. Sie enthalten namentlich die gesuchten Aufklärungsobjekte (Namen von Personen, Organisationen oder Unternehmen, Adressteile usw.) sowie die Liste der Schlüsselwörter (Key Words), von denen der Auftraggeber erwartet, dass sie in den abgehörten Kommunikationen erscheinen. All diese Informationen sind zur Ausarbeitung automatischer Filtersysteme für die Kommunikationen notwendig. Je nach Auftrag können zwischen fünf und mehrere hundert Schlüsselwörter eingegeben werden. Im Bereich der Bekämpfung der Proliferation beispielsweise zählt die Liste der Schlüsselwörter mehr als zehn Seiten mit 25 Begriffen pro Seite. [...] "Diese Aufträge beziehen sich auf die Bekämpfung der Proliferation, die Spionageabwehr, das organisierte Verbrechen, den Kampf gegen den Terrorismus und die Situation im Golf." [...] "Die Filtrierung erfolgt mit Hilfe von Systemen künstlicher Intelligenz. Diese Systeme vergleichen den Inhalt der Kommunikation mit den vordefinierten Adressierungselementen und Schlüsselwörtern." [...] "Meldungen, die keinen dieser Kriterien entsprechen, werden automatisch herausgefiltert."*

Ferner ist es auch möglich, formale Selektoren einzusetzen: besteht ein konkretes *"Adressierungselement"*[60] kann sämtliche Kommunikationen eines bestimmten Teilnehmers abgefangen werden und – falls die Datenmengen zu gross sind – zusätzlich auf inhaltliche Suchbegriffe zurückgegriffen werden. [30, S.1517ff.]

Rechtlich bemängelt die Geheimdienstkontrolle die Intransparenz der massenweisen Funküberwachung, weil sie es den Bürgern erschwere *"mit einem adäquaten Verhalten reagieren zu können"*[61] und weist zudem darauf hin, dass es zu illegalen Aufträgen seitens des damali-

[58] Dem zuvor standen allerdings Medienberichte, die bis 1999 zurückreichen.

[59] Schweizer National- und Ständerat

[60] Im Beispiel eine Telefax-Nummer.

[61] Hiermit sagt die *GPDel* implizit, dass die Funk-Massenüberwachung unter Bewusstsein der Bevölkerung, einen *Chilling Effect* – durch angepasstes Verhalten – zur Folge haben müsste.

gen Inlandsgeheimdienstes *DAP*[62] kommt. Dieser reagiert offenbar mit der Aussage darauf, die Rechtsgrundlagen müssten *"korrigiert"* werden. Auch der Schweizer Bundesrat[63] nimmt eine fragwürdige Rolle ein, indem er die illegalen *DAP*-Aufträge mittels Verordnung bewilligt – ohne dafür eine gesetzliche Grundlage zu haben. Es ist die Rede davon, dass der Bundesrat entschieden habe, *"die Normenhierarchie umzukehren"*. [30, S.1519,1523ff.]

Betreffend die Legitimität zur Auslandsspionage stellt die *GPDel* im Wesentlichen fest, dass im Sinne der Weltraumtheorie, die Funk-Massenüberwachung als unproblematisch betrachtet werden kann, weil der Weltraum internationales Gemeingut und somit rechtsfrei sei. Zudem stellt sie fest, dass keine Übereinkommen existieren, welche eine Auslandsspionage verbieten würden. Grundrechte machten an Staatsgrenzen halt – die Sicherheitsinteressen des Staates gingen vor. In wirtschaftlicher Hinsicht zudem könnten Daten aus der Auslandsspionage situationsbedingt nützliches Tauschmittel sein. Allerdings wird betont, dass die Bedeutung der Funk-Massenüberwachung abnehme,[64] weil mehr Kommunikation über Kabel verlaufen würde.[65] Abschliessend werden die Finanzierung des *Onyx*-Systems und die zunehmende Verschlüsselung der Kommunikation als Herausforderungen gesehen.[66] [30, S.1527ff.,1531ff.]

Spätere Berichte der *GPDel* informieren – bis heute – nicht weiter über die Funkaufklärung: bloss 2007 und 2008 sind Notizen zu vernehmen, wonach *"Wirksamkeitskontrollen"* (weiterhin) ausblieben. [32, S.5149][33, S.2624]

2.4.4.2 Geplante Kabelaufklärung im NDG

Am 25. September 2015 hat das Schweizer Parlament beschlossen, das *Nachrichtendienstgesetz* einzuführen. Dieses soll einerseits die zwei bisherigen Gesetzestexte *BWIS* und *ZNDG*, die die Arbeit des zivilen Geheimdienstes NDB regeln, formell-gesetzlich einigen und damit obsolet machen, führt gleichzeitig aber einen ganzen Katalog an neuen Überwachungsmassnahmen ein, die die Machtfülle des NDB erheblich erweitern: eine davon ist die "Kabelaufklärung".

In Art. 39 NDG wird das grundsätzliche Wesen der Kabelaufklärung dargelegt:

1. *Der NDB kann den durchführenden Dienst damit beauftragen, zur Beschaffung von Informationen über sicherheitspolitisch bedeutsame Vorgänge im Ausland (Art. 6*

[62] Der Dienst für Analyse und Prävention *DAP* war bevor er mit dem Auslandsgeheimdienst *SND* zusammengelegt wurde, jener Schweizer Geheimdienst, der strikte für das Inland zuständig war. Weil das *Onyx*-Funküberwachungssystem eigentlich für die Massenüberwachung des Auslands bestimmt war, hat die *GPDel* durch die schon 2003 existierenden DAP-Aufträge an das *Onyx*-System illegales Handeln festgestellt.

[63] Die Bundesregierung der Schweiz

[64] Es wird festgestellt, dass nur 1% der Telefongespräche über Satellit abgewickelt würden.

[65] Insbesondere schlecht vernetzte Gebiete sind auf satellitengestützte Kommunikation angewiesen.

[66] Dies ist Indiz dafür, dass durch den massenweisen Einsatz von Verschlüsselungstechnik die Kosten der Massenüberwachung derart erhöht werden können, dass die Finanzierung solcher Praxen erschwert wird.

Abs. 1 Bst. b) sowie zur Wahrung weiterer wichtiger Landesinteressen nach Artikel 3 grenzüberschreitende Signale aus leitungsgebundenen Netzen zu erfassen.

2. *Befindet sich sowohl der Sender als auch der Empfänger in der Schweiz, so ist die Verwendung der erfassten Signale nach Absatz 1 nicht zulässig. Kann der durchführende Dienst solche Signale nicht bereits bei der Erfassung ausscheiden, so sind die beschafften Daten zu vernichten, sobald erkannt wird, dass sie von solchen Signalen stammen.*

3. *Daten aus erfassten Signalen dürfen nur an den NDB weitergeleitet werden, wenn deren Inhalt den für die Erfüllung des Auftrags definierten Suchbegriffen entspricht. Die Suchbegriffe sind so zu definieren, dass ihre Anwendung möglichst geringe Eingriffe in die Privatsphäre von Personen verursacht. Angaben über schweizerische natürliche oder juristische Personen sind als Suchbegriffe nicht zulässig.*

In den Abs. 1 und 2 wird suggeriert, einzig Daten, die ins Ausland fliessen oder vom Ausland in die Schweiz gesendet werden, seien von der neuen Möglichkeit zur Massenüberwachung betroffen.

Technisch muss entgegengehalten werden, dass einerseits zahlreiche Dienste, welche die Schweizer Bevölkerung nutzt, im Ausland betrieben werden – darunter auch viele Schweizer Online-Medien oder E-Mail-Dienste, was bedeutet, dass der Datenverkehr zwangsläufig *"grenzüberschreitend"* ist; andererseits kann kein gewöhnlicher Internetteilnehmer entscheiden, über welche Wege die eigene Kommunikation im Internet geleitet wird.[67] Somit ist möglich, dass eine versendete E-Mail, die beispielhaft eine Kommunikation innerhalb der Schweiz konstituiert, über den Frankfurter Internet Exchange Point und die USA verkehrt, was heisst, dass nicht nur der deutsche Auslandsgeheimdienst *BND* die Nachricht abfangen und verwerten kann, sondern auch der britische Auslandsgeheimdienst *GCHQ* im Rahmen seines *Tempora*-Programms und rein rechtlich auch der NDB: sowohl (zumindest) einmal aus- als auch eingangs.

Die einzige "echte" Einschränkung, die in Art. 39 mit Abs. 3 sichtbar ist, wäre die, dass zumindest unmittelbar keine Named Entitites natürlicher (etwa: Personennamen) oder juristischer Personen (etwa: Firmennamen) als Selektoren erlaubt sind: es gilt allerdings zu beachten, dass damit nicht ausgeschlossen ist, dass diese Begriffe in der Kommunikation auftauchen. Es ist nur nicht erlaubt, sie direkt als Suchbegriffe einzusetzen. Gelingt es also, Begriffe zu finden, welche in anderen (und gegebenenfalls mehr) Worten diese Named Entities im Kern referenzieren, wie dies beispielsweise bei Topikalischer Analyse mittels Verfahren wie Latent Dirichlet Allocation [8] gelingen kann, so ist auch diese Regelung technisch ausgehebelt.

[67] Es wird geschätzt, dass 60% des Schweizer Internetverkehrs über das Ausland abgewickelt wird; $\frac{1}{3}$ des Schweizer Internetverkehrs soll sogar über Grossbritannien laufen, was ihn *XKeyscore* sicher zugänglich macht. [53, 71]

Für die in Art. 39 Abs. 2 festgehaltene Regel, dass Kommunikation zwischen Entitäten im Inland zu löschen sei, muss einerseits festgehalten werden, dass technisch die Kontrolle darüber für die Öffentlichkeit oder die Betroffenen[68] nicht möglich ist und andererseits bestehen normative Hintertüren. Art. 42 Abs. 2 und 3 führen nämlich folgende Regelungen:

- *Er leitet ausschliesslich Daten an den NDB weiter, die Informationen zu den für die Erfüllung des Auftrags definierten Suchbegriffen enthalten. Informationen über Personen im Inland leitet er nur dann an den NDB weiter, wenn sie für das Verständnis eines Vorgangs im Ausland notwendig sind und zuvor anonymisiert wurden.*
- *Enthalten die Daten Informationen über Vorgänge im In- oder Ausland, die auf eine konkrete Bedrohung der inneren Sicherheit nach Artikel 6 Absatz 1 Nachrichtendienstgesetz stabe a hinweisen, so leitet der durchführende Dienst sie unverändert an den NDB weiter.*

Es ist nicht deutlich, was Art. 42. Abs. 2 mit *"Informationen über Personen im Inland"* genau meint: auf jeden Fall kann der NDB die unveränderten Rohdaten einsehen, wenn er sich nach Abs. 3 auf die *"innere Sicherheit"* beruft.

Auch die Schweizer Geheimdienstkontrolle sieht in der Kabelaufklärung gemäss einem Bericht wörtlich die Gefahr, dass sie zur *"Suche im Heuhaufen"* ausufern könnte. [31, S.28]

2.5 Befunde: Konkrete Selektoren zur Massenüberwachung

An dieser Stelle werden Selektoren verschiedener Art aufgezeigt, die im Rahmen der Snowden-Enthüllungen konkretisiert werden konnten.

2.5.1 FVEY: Mögliche Echelon-Selektorenlisten

Im Web finden sich Listen von Wörtern oder Ausdrücken, die mit dem *Echelon*-System in Verbindung gebracht werden: zur Illustration hat *The Register* eine solche im Zuge der um die Jahrtausendwende stattfindenden Diskussion über die globalisierte Überwachung veröffentlicht. [46]

Eine Anfrage[69] im Rahmen der Masterarbeit gegenüber Duncan Campbell, der das *Echelon*-System schon früh enthüllt [14] hat, ergab, dass die Liste eine Erfindung darstellt und (entsprechend) keine Rückschlüsse über die Funktionsweise der Selektorgenerierung ermöglicht.

[68] Diese erhalten in aller Regel davon nicht einmal Kenntnis, sondern können nur diffus erahnen, dass sie überwacht werden.

[69] Im November 2015 erfolgt.

2.5.2 FVEY: XKeyscore und Beispiele von Selektoren

Konkrete Evidenz von Soft-Selektoren finden sich im Zusammenhang mit *XKeyscore*-Veröffentlichungen. Wie das *Linux Journal* erstaunt [94] feststellt, werden deren Besucher, sofern sie nicht aus den *FVEY*-Staaten stammen, als "Extremisten" markiert, die weitergehend zu überwachen sind. Gleichsam werden Besucher der Webseite der Tails-Linux-Distribution kategorisiert: diese Konfiguration von Anwendungssoftware auf Linux-Basis erlaubt spurenarmes Surfen und ist auf anonymisierende und verschlüsselte Kommunikation bedacht.

Gemäss von Edward Snowden den Medien gegenüber geleaktem *XKeyscore*-Code im Anhang[70] unter C.1 macht sich als "Extremisten" ferner verdächtig, wer Begriffe der Selektorenmenge {`tails`, `Amnesiac Incognito Live System`} in einer Kombination mit Begriffen der Selektorenmenge {`USB`, `CD`, `secure desktop`, `IRC`, `truecrypt`, `tor`} in seinem Datenstrom führt.

Es sind zudem in *XKS*-Folien viele Beispiele von möglichen Selektoren gegeben. Ein gutes Beispiel ist für den Bereich des Advanced Conventional Weapons *ACW*-Waffenhandels vorhanden [81], wo diverse Selektormengen bestehen, die mittels den logischen Operatoren AND und ODER miteinander verknüpft werden können. Konkret besteht ein "Example 4", das zunächst fünf Selektormengen aus Einzelworten oder Wort-n-Grammen führt:

$$S_{acwitems} = \{\texttt{machine gun, grenade, AK 47}\} \tag{2.1}$$

$$S_{acwpositions} = \{\texttt{minister of defence, defense minister}\} \tag{2.2}$$

$$S_{acwcountries} = \{\texttt{somalia, liberia, sudan}\} \tag{2.3}$$

$$S_{acwbrokers} = \{\texttt{south africa, serbia, bulgaria}\} \tag{2.4}$$

$$S_{acwports} = \{\texttt{rangood, albasra, dar es salam}\} \tag{2.5}$$

Damit kann eine Suchabfrage erfolgen, die beispielsweise – wie vorgeschlagen – darin besteht, einen oder mehrere der Selektoren in den Selektormengen $S_{acwitems}$ und $S_{acwpositions}$ (AND-verknüpft) mit einem beliebigen Begriff einer der anderen drei (ODER-verknüpften) Selektormengen zu verbinden.

Zwei Beispiele gültig kombinierter Selektoren, mit denen jeweils AND-verknüpft in Datenströmen gesucht werden könnte, wären:

$$S_1 = \{\texttt{machine gun, defense minister, liberia}\} \tag{2.6}$$

$$S_2 = \{\texttt{AK 47, grenade, minister of defence, south africa, rangood}\} \tag{2.7}$$

[70]Vgl. Z.91ff.

2.5.3 USA: WikiLeaks-Publikationen

Mitte 2015 hat *WikiLeaks* damit begonnen, beispielhaft einige Selektoren zu veröffentlichen, die von ihrer Natur her als formale, oder Strong-Selektoren zu qualifizieren sind.

Dabei bestehen die bisher veröffentlichten Selektoren, die die Länder Deutschland, Frankreich und Japan direkt betreffen, jeweils aus Telefonnummern der Regierungsspitzen oder von Akteuren der Industrie: mit Japan sind beispielsweise Stellen von *Mitsubishi* oder *Mitsu* als Selektoren definiert. [125, 126, 127]

2.5.4 Deutschland: Medienberichte und NSAUA

Anfangs Mai 2015 wurde bekannt [68], dass der *BND* rund 12'000 NSA-Selektoren gelöscht hat, die folgende Teilmenge an Begriffen enthalten hat: {`.eu`, `Bundesamt`, `diplo`, `gov`}. Offensichtlich wurden solche Selektoren dazu genutzt, amtlichen (deutschen und EU-institutionellen) Datenverkehr abzufangen.

Es kann davon ausgegangen werden, dass mehr über konkrete vom *BND* eingesetzte (*NSA*-)Selektoren herausgefunden wird, denn seit Ende November 2015 – mit Abschluss dieser Arbeit – sind Mitglieder des *NSAUA* erstmals dabei, selber Einblick in (bestimmte) Selektorenlisten zu nehmen, die beim Bundeskanzleramt lagern. [95]

2.5.5 Schweiz: Medienberichte und Onyx

Konkrete Suchbegriffe von *Onyx* sind öffentlich nicht dokumentiert. Im Artikel der Weltwoche von 2005 wird allerdings ein konkretes Beispiel angeführt [22], das sich auf eine Expertenmeinung abstützt und Beispielen für den Bereich Waffenhandel nahe kommt, die in *XKS*-Folien geschildert werden und zudem thematischen Charakter haben:

> *Wie die Internetsuchmaschine Google innert Sekunden das endlos scheinende WWW-Meer nach den gewünschten Begriffen ausfischt, so kann auch Onyx den gesamten Telefon-, Fax- und Mailverkehr, der über Satelliten läuft, permanent und methodisch clever überwachen. Je nach Auftrag werden zwischen fünf und mehreren hundert Begriffen eingegeben. Je präziser die Schlüsselwörter, desto exakter die Resultate. Allgemeine Ausdrücke wie "Terrorismus", "Bombe" oder "Anthrax" sind laut Spezialisten ungeeignet. Die Verknüpfung konkreter Städtenamen wie "Riad", "Bagdad" oder "Falludscha" mit Sprengstoffen wie "TNT", "Anfos" oder "RDX" und den Namen verdächtiger extremer Muslime im Mittleren und Nahen Osten hingegen ist als Filter bereits durchaus geeignet. Nach dem gleichen Muster werden derzeit konkret auch Vorgänge in der russischen Politik und Wirtschaft (vom Handel mit Erdgas bis zum Verkauf von radioaktivem Material), in Transkaukasien und auf dem indischen Subkontinent ausgehorcht.*

Hierbei werden Named Entities aus dem Waffenbereich mit solchen (dazu passender) geografischer Ortschaften verknüpft.

3 Automatische Generierung von Selektoren

Der Vergleich mit Snowden [...] ist völlig daneben, weil die NSA, so, wie wir das heute wissen, einfach wirklich alles sammelt und auf die Seite legt. Bei uns suchen wir nach der Nadel im Heuhaufen, wir legen Kriterien fest, zeitlich befristet, und schauen, ob wir dann etwas finden. Alles, was nicht dem entspricht, wird ohnehin gar nicht verwertet. Wir haben am Schluss nur ganz weniges; es ist nicht einfach eine Sammelübung. Sie können das auch beim Personalbestand sehen. Wenn Sie vergleichen, wie viele Leute sich bei der NSA und wie viele Leute sich in unserem Nachrichtendienst damit beschäftigen, dann stellen Sie fest, dass wir im Bereich von 0,01 Promille sind.

– Bundesrat Ueli Maurer am 11. Juni 2015 im Schweizer Ständerat [120]

Im Rahmen dieses Kapitels werden nach drei unterschiedlichen Modellen Soft-Selektoren generiert, die dann im Rahmen des Kapitels 4 evaluiert werden.

3.1 Grundannahmen

Die Überwachungsmethodologie, die nachgebildet wird, entspricht bildlich der Suche nach einer Nadel im Heuhaufen, im Englischen auch *needle-in-a-haystack* genannt.

Folgende Grundannahmen werden getroffen und bilden die Grundlage des eigenen Überwachungssettings, das rein exemplarischen Charakter hat, ab:

- Aller Text eines Datenstroms oder einer Textkollektion ist grundsätzlich dem Generalverdacht ausgesetzt, verdächtiger Text nach einer bestimmten Kategorie zu sein. Eine Unterscheidung nach Diensten oder konkretem Textinhalt unterbleibt.
- Aller Text wird durchsucht, um Verdacht schöpfen zu können: dies qualifiziert die Methodologie als eine der Massenüberwachung, weil unabhängig von einem Verdacht und ohne Anlass, in möglichst vielen Daten nach Verdachtsmomenten gesucht wird.
- Als Texteinheit gelten Dokumente, wie sie von Websuchmaschinen indexiert werden: das sind beispielsweise HTML-Seiten, PDF-Dokumente, Bilder oder Dateien der Tabellenkalkulation; prinzipiell können das alle Dateien sein, die sich textuell durchsuchen lassen.
- Es wird angenommen, dass Dokumente dieser Art auch in den Datenströmen von

gewöhnlichen Internet-Usern auftreten können.

- Es wird nicht davon ausgegangen, dass Verdächtige konkret schon bekannt sind: diese sollen erst gefunden werden. Das heisst gleichzeitig, dass keine zusätzliche Suche nach konkreten Telekommunikationsmerkmalen oder Metadaten vorgenommen werden soll, wie sie mit formalen oder Strong-Selektoren vollzogen würde. Stattdessen kommen rein inhaltliche, oder "Soft"-Selektoren zum Einsatz.
- Es muss eine Vorstellung davon entwickelt werden, was verdächtiges Textmaterial an der Oberfläche auszeichnet, um textuell durchsuchbare Daten einer Auswahl verdächtigen Materials zuzuführen.
- Es wird angenommen, dass sich noch unbekannte Verdächtige gleich oder ähnlich ausdrücken wie bereits bekannte Verdächtige.
- Soft-Selektoren werden folglich aus Textmaterial generiert, das bekannten Verdächtigen zugesprochen wird.
- Selektoren in diesem Szenario sind Wörter (einzeln, in bestimmter Wortreihenfolge oder Kombination), welche zur Suche nach verdächtigen Texten genutzt werden.

3.2 Trainingsdaten

Diese Sektion beschreibt die Auswahlkriterien von zwei Trainingskorpora, die zur Generierung der Soft-Selektoren genutzt werden, sowie die dafür nötigen Aufbereitungsschritte des Trainingsmaterials.

3.2.1 Auswahlkriterien

Der Schweizer Geheimdienst NDB führt alljährlich Berichte über die "Sicherheit Schweiz" durch, worin über die Jahre sichtbar [74, 75] ist, dass namentlich zwei Gruppen auffallen, welche die "links-", respektive "rechtsextreme" Szene dominieren:

- Revolutionärer Aufbau Schweiz RAS[71] (im Folgenden *Aufbau*) mit einer Webpräsenz unter `http://www.aufbau.org/`.[72]
- Partei National Orientierter Schweizer PNOS[73] (im Folgenden *PNOS*) mit einer Webpräsenz unter `http://www.pnos.ch/`.[74]

[71]Es gibt auch eine (starke) Zürcher Sektion, die RAZ abgekürzt wird.

[72]Abruf: 28. November 2015.

[73]Es gibt auch diverse kantonale Sektionen.

[74]Abruf: 28. November 2015.

Die zwei gewählten Gruppen haben den Vorteil, dass sie medial offen in Erscheinung treten, so dass textuelles Webmaterial vorhanden ist, das als Trainingsmaterial genutzt werden kann, um (1) zumindest einzelne Texte, (2) womöglich ähnlich gesinnte Sub- oder Teil-Gruppen oder (3) (gänzlich) andere Gruppen zu finden, die wesentliche Aspekte ihres jeweiligen “politischen Extremismus” teilen.[75] Die Suche danach erfolgt unter Einsatz daraus generierter Soft-Selektoren, die zur Rasterung von Textkollektionen (von Datenströmen) genutzt wird. Welche Daten das in der Wirklichkeit sein können, hängt von den Möglichkeiten ab, die Geheimdiensten zur Verfügung stehen. Weiteres hierzu ist im Kapitel 4 zur Evaluation Thema.

3.2.2 Aufbereitung

Jedes Trainingskorpus ist in folgenden Schritten bezogen und aufbereitet worden, wobei Duplikate mittels *fdupes(1)* nach allen wesentlichen Schritten stets entfernt wurden:

1. Die Webseiten werden mittels *wget(1)* heruntergeladen.
2. Die Verzeichnisstrukturen werden aufgehoben: alle Dateien werden flach (nebeneinander) gespeichert.
3. Duplikate werden durch *fdupes(1)* entfernt.
4. Mittels *Apache Tika* wird von allen Dateien[76] der textuelle Inhalte extrahiert.
5. Bestehendes, gleichbleibendes Navigationsmaterial am Anfang und Ende von Dokumenten (oder andere Textartefakte) werden mittels eines eigenen Skripts entfernt.
6. Die Wörter werden normalisiert (Kleinschreibung ohne Umlaute).
7. Stoppwörter werden in den drei Landessprachen Deutsch, Französisch und Italienisch entfernt; auch englische Stoppwörter werden entfernt.
8. Mittels *Apache Tika* werden diejenigen Texte identifiziert und behalten, die (mehrheitlich) deutschsprachig sind.
9. Die Dateien werden tokenisiert: die Dokumente bestehen fortan aus rein textuellen Dateien mit einem Token je Zeile.

3.2.3 Die Trainingskorpora Aufbau und PNOS

Die *Aufbau*- und *PNOS*-Korpora entstammen den Webseiten aufbau.org und pnos.ch mit Stand vom 10. Oktober 2015 und weisen die in Tabelle 3.1 aufgezeigten Merkmale deskrip-

[75]Dies als Möglichkeiten davon, was im Spektrum der True-Positives möglicher Treffer liegt.

[76]Das sind auch beispielsweise PDF-Dateien.

tiver Statistik (im Vergleich) auf.[77]

Merkmal	*Aufbau*	*PNOS*
Korpusgrösse (roh)	161MB	15MB
Korpusgrösse (roh; duplikatfrei)	154MB	15MB
Anzahl Dateien (roh; duplikatfrei)	3′282	727
Anzahl Dateien (rein textuell)	1′848	676
Anzahl Dateien (rein textuell; ohne Textartefakte)	1′478	672
Anzahl Dateien (rein textuell; deutschsprachig)	1′094	574
Anzahl Token	361′748	154′699
Anzahl Types (Token-to-Type-Ratio)	38′405 (10.62%)	25′835 (16.7%)

Table 3.1: Deskriptive Statistik: Korpora *Aufbau* und *PNOS*

3.3 Methode 1: TFIDF-Modell

Beim TFIDF-Modell soll auf Basis relativ häufiger und gut über die Dokumente verteilter, aber nicht häufigster und überall (in allen Dokumenten) vorkommender Selektoren, die keine Stoppwörter enthalten, überwacht werden. Dies geschieht mit drei Arten von Selektoren:

- Einzelworte
- Wort-2-Gramme
- Wortkombinationen (von 2–5 Worten)

Für jeden der beiden Korpora *Aufbau* und *PNOS* werden die fünf besten Selektoren pro Art ausgewählt, so dass auf Basis jeden Korpus' 15 Selektoren entstehen: zusammen werden in diesem Modell 30 Selektoren generiert.

3.3.1 Grundlagen

Das TFIDF-Modell stellt das klassische Modell[78] des *Information Retrieval* dar, wobei erstens eine Funktion tf besteht, um die Termfrequenz je Term und Dokument zu fassen. Zweitens besteht eine Funktion idf, um die Bedeutung eines Terms im Korpus[79] zu er-

[77] Weiter unten in der Tabelle stehenden Ergebnisse sind jeweils auf die vorhergehenden Ergebnisse aufgebaut, das heisst wurden den dafür erforderlichen Prozessierungsschritten genauso unterworfen. Beispielsweise ist die Anzahl der Dateien, wo deutschsprachige Dokumente gemeint sind, eben auch frei von Textartefakten.

[78] Vgl. beispielsweise die deutsche Wikipedia vom 16. Oktober 2015 hierzu. URL `https://de.wikipedia.org/w/index.php?title=Tf-idf-Ma%C3%9F&oldid=147045152`. Abruf: 26. November 2015.

[79] Über alle Dokumente hinweg betrachtet.

messen, die wie folgt definiert ist:

$$idf_t(t, D) = log\frac{N}{n_t} \tag{3.1}$$

d bezeichnet ein konkretes Dokument, t ein Token, wobei D die Menge aller Dokumente darstellt (mit N als deren Mächtigkeit oder Anzahl der Dokumente).

Das Produkt von tf und $tfidf$

$$tfidf(d, t, D) = tf(d, t) * idf_t(t, D) \tag{3.2}$$

stellt den $tfidf$-Wert dar, welcher jene Terme hoch bewertet, die zwar häufig vorkommen (tf-Wert), dies aber nur, wenn der Term in der Kollektion der Dokumente ebenfalls als bedeutend gilt (durch den idf-Wert). $tfidf$-Werte gelten immer für ein bestimmtes token t in Bezug zu einem Dokument d in einer Textkollektion D.

3.3.2 Einzelworte

Die einfachste Ausführung des TFIDF-Modells wählt je Korpus die fünf Einzelworte, die den höchsten TFIDF-Wert haben, als Selektoren aus.[80]

Beispiele von Selektoren hierbei sind {`prozess`} oder {`frauen`} beim *Aufbau*- und {`rütli`} oder {`schweizer`} beim *PNOS*-Korpus.[81]

3.3.3 Wort-2-Gramme

Als Selektoren werden je Korpus fünf Wort-2-Gramme gesucht, wie sie innerhalb der einzelnen Dokumente vorkommen: im Rahmen des *Aufbau*-Korpus entstehen dabei 168'992 und beim *PNOS*-Korpus 92'326 Auswahlmöglichkeiten für Selektoren.

Beispiele resultierender Selektoren[82] sind schliesslich ⟨`antirassistische,aktion`⟩ (Korpus: *Aufbau*) oder ⟨`tobias,hirschi`⟩[83] (Korpus: *PNOS*).

3.3.4 Wortkombinationen von 2–5 Worten

Bei diesem Selektionsverfahren werden alle Wortkombinationen – ohne Wortwiederholungen – von zwei bis fünf Worten berücksichtigt, was mathematisch Mengen entspricht, die in

[80] Zum genauen Auswahlverfahren, das durchgehend beim Überwachungsmodell 1 und 2 zur Anwendung kommt, siehe die Sektion 3.6.

[81] Weitere Selektoren nach dem Modell der TFIDF-Einzelworte sind Tabelle A.1 zu entnehmen.

[82] Die restlichen Selektoren sind dem Anhang in Tabelle A.2 zu entnehmen.

[83] Als Named Entity aufgefasst handelt es sich hierbei um einen bekannten Akteur der PNOS-Gruppierung.

jeder möglichen Kombination entweder zwei, drei, vier oder fünf Elemente umfassen; die maximale Anzahl solcher Mengen beträgt entsprechend:

$$\sum_{i=2}^{5} \frac{n!}{(n-i)!i!} \tag{3.3}$$

Im theoretischen Beispiel: Wenn im Minimum $n = 5$ ist, oder in der gesamten Textkollektion hypothetisch wenigstens fünf distinkte Worte (Types) vorhanden sind, gegeben in $T = \{t_1, t_2, t_3, t_4, t_5\}$, so lassen sich daraus alleine für 2-Wort-Kombinationen folgende zehn Selektoren S_1 bis S_{10} konstruieren:

$$S_1 = \{t_1, t_2\} \tag{3.4}$$

$$S_2 = \{t_1, t_3\} \tag{3.5}$$

$$S_3 = \{t_1, t_4\} \tag{3.6}$$

$$S_4 = \{t_1, t_5\} \tag{3.7}$$

$$S_5 = \{t_2, t_3\} \tag{3.8}$$

$$S_6 = \{t_2, t_4\} \tag{3.9}$$

$$S_7 = \{t_2, t_5\} \tag{3.10}$$

$$S_8 = \{t_3, t_4\} \tag{3.11}$$

$$S_9 = \{t_3, t_5\} \tag{3.12}$$

$$S_{10} = \{t_4, t_5\} \tag{3.13}$$

Hinzu kommen weitere zehn Selektormengen S_{11} bis S_{20} als 3-Wort-Kombinationen, deren fünf als 4-Wort-Kombinationen (S_{21} bis S_{25}) und zuletzt der Selektor als Selektormenge $S_{26} = T$, der alle Elemente t_1 bis t_5 fasst: in der Summe also 26 Selektoren für einen Korpus, der aus nur fünf Worten besteht. Ist die Textkollektion sogar aus nur einem Dokument konstituiert, so wird klar, dass es mit unserem einfachen Setting 26 verschiedene Arten gibt, dieses eine Dokument zu "filtern".

Es entstehen aus dem *Aufbau*-Korpus Selektoren[84] wie {`8.`, `märz`}, {`abdallah`, `ibrahim`}; oder {`1.`, `august`} und {`heinz`, `kaiser`} aus dem *PNOS*-Korpus.

3.4 Methode 2: Verdachtssprache-Modell

Beim korpuslinguistisch motivierten Verdachtssprache-Modell wird entsprechend den Befunden von Ebling et al. [21] davon ausgegangen, dass es sprachliche Indikatoren dafür

[84] Weitere Selektoren sind im Anhang mit Tabelle A.3 sichtbar.

geben kann, dass in einem Text “politischer Extremismus” vorliegt.

Es werden dabei in den zwei Trainingskorpora Wortkombationen gesucht, die nach dem gleichen Verfahren wie beim TFIDF-Modell ausgesucht werden, allerdings mindestens ein Inhaltswort als Indikator enthalten müssen, dass die Ausdrücke als intensivierend,[85] skandalisierend oder verschwörend erscheinen lässt.

Bei dieser Methode entstehen für jedes Trainingskorpus 30 Selektoren, die Wortkombinationen von zwei bis fünf Worten der folgenden Art repräsentieren:

- intensivierend (generell)
- intensivierend (absolut)
- intensivierend (hoch)
- intensivierend (extrem hoch)
- skandalisierend
- mit Verschwörungsvokabular

3.4.1 Grundlagen

“Politischer Extremismus” wird gemäss Ebling et al. [21] in vier Dimensionen angenommen:

1. *Ablehnung des demokratischen Verfassungsstaates*
2. *Dogmatismus und Commitment*
3. *Verschwörungstheorien*
4. *Fanatismus: Bereitschaft zur gewaltsamen Propagierung und Durchsetzung der erstreben Ziele*

Es wird davon ausgegangen, dass eine Häufung von eindeutig skandalisierenden Begriffen, relativierenden oder intensivierenden Ausdrücken oder anderen Gradpartikeln “politischen Extremismus” anzeigen kann.

Obwohl auch Wortreihenfolgen[86] oder -markierungen eine wesentliche Rolle spielen können, um beispielsweise Ausdrucksrelativierungen zu erkennen,[87] wird im Rahmen dieser Arbeit rein auf Wortkombinationen von zwei bis fünf Worten gesetzt, die mindestens eines der

[85] Hierbei in vier verschiedenen Stufen: generell, absolut, hoch und extrem hoch.

[86] Im Sinne der Bildung von Wort-n-Grammen.

[87] Zu nennen sind Beispiele wie ein Wort – etwa: Freiheit – in Anführungszeichen zu setzen oder von der “sogenannten Demokratie” zu sprechen. Das kann anzeigen, dass die bestehende Verfassungsordnung abgelehnt wird.

Wörter gemäss Tabelle 3.2 enthalten.[88]

Wortkategorie	Beispiele	Anzahl Wörter
Intensivierer (generell)	`allererst, durchaus, ekelhaft, schlichtweg, schrecklich`	83
Intensivierer (absolut)	`absolut, entschieden, komplett, schlichtweg, zweifellos`	39
Intensivierer (hoch)	`besonders, gerade, merklich, solch, wesentlich`	22
Intensivierer (extrem hoch)	`allerbest, ekelhaft, mega, traumhaft, unsaeglich`	40
Skandalisierer	`begriffsstutzig, frivol, kindisch, schwindel, unwahrhaftig`	892
Verschwörungsvokabular	`eigentlich, entlarven, hinterzimmer, vorgaukeln, vormachen`	80

Table 3.2: Verdachtssprache-Modell: Intensivierer, Skandalisierer und Verschwörungsvokabular

3.4.2 Intensivierende Wortkombinationen

Die Anzahl möglicher Selektoren beträgt für intensivierende Wortkombinationen vom Typ "generell", "absolut", "hoch" und "extrem hoch" 10'618, 8'979, 2'905 und 2'601, was den *Aufbau*-Korpus betrifft.

Beispiele von Selektoren sind:

- generell: {`kriminalisiert`, `repression`, `stark`}
- absolut: {`fluchthilfe`, `leisten`, `praktisch`}
- hoch: {`kampf`, `mehrfach`, `stark`}
- extrem hoch: {`flüchtlingsaufnahme`, `möglichst`, `unattraktiv`}

Beim *PNOS*-Korpus stehen für die Wortkombinationen vom intensivierenden Typ "generell", "absolut", "hoch" und "extrem hoch" 8'821, 6'842, 2'622 und 1'292 Selektoren zur Auswahl.

Nach der Auswahl nach TFIDF-Mass bester Selektoren verbleiben beispielhaft folgende Selektoren:

- generell: {`meinungsfreiheit`, `punkten`, `stark`}
- absolut: {`nationalen`, `probleme`, `rein`}
- hoch: {`allzu`, `stark`, `systemeliten`}

[88] Die Beispiele in der Tabelle stellen je fünf Wörter dar, die zufällig ausgewählt wurden und alphabetisch sortiert sind.

- extrem hoch: {einheit, völkische, weitestgehend}

3.4.3 Skandalisierende Wortkombinationen

Bei den skandalisierenden Kombinationen entstehen beim *Aufbau*-Korpus 6'254 und beim *PNOS*-Korpus 4'808 verschiedene Selektoren.

Beispiele davon sind:

- *Aufbau*: {angegriffen, grausam, unzählige, zivilistinnen}
- *PNOS*: {261, nutzlos, volkes}

3.4.4 Wortkombinationen mit Verschwörungsvokabular

Die Anzahl Wortkombinationen mit Verschwörungsvokabular betragen im *Aufbau*-Korpus 5'580 und im *PNOS*-Korpus 2'828.

Zu den besseren Selektoren zählen nach eingesetztem TFIDF-Mass die Folgenden:

- *Aufbau*: {angeklagten, dienen, eigentlich}
- *PNOS*: {führen, gegner, lügen}

3.5 Methode 3: LDA-Modell

Die Methode der topikalischen Analyse setzt darauf, mit relativ vielen – thematisch zusammenhängenden – Wörtern, Treffer zu erzeugen.

3.5.1 Grundlagen

Bei Latent Dirichlet Allocation *LDA* handelt es sich um ein Verfahren des Maschinellen Lernens *ML* – konkreter des Topic Modeling *TM*, das dazu dient aus einer Kollektion von Texten Topics (oder Themen) zu erstellen, die selber wiederum im Sinne eines generativen Ansatzes, die Textkollektion repräsentieren. Topics sind eine Menge von Wörtern, die in ihrem Zusammenhang den Anspruch erheben, ein – latentes, zwischen den Wörter liegendes – semantisches Themenfeld der gesamten Textkollektion zu repräsentieren. Von der Inspiration her, lehnt es sich an bereits früher dargelegte Verfahren der Latent Semantic Analysis[89] *LSA* an, hat allerdings neben der linear-algebraischen auch eine stochastische

[89]Auch: Latent Semantic Indexing *LSI*

Basis, die bei den zumindest ursprünglichen Verfahren[90] von *LSA* fehlt.[91] Das Verfahren wurde von Blei 2003 [8] und mathematisch ausführlich beschrieben. Die mathematischen Grundlagen nur oberflächlich wieder zu geben, werden dem komplexen Verfahren allerdings einerseits nicht gerecht, andererseits hat diese Arbeit auch nicht den Schwerpunkt darauf, auf Verfahren des *TM* im Detail einzugehen, so dass – mangels Möglichkeit der mathematischen Komplexität im Rahmen dieser Masterarbeit gerecht zu werden – darauf verzichtet wird.

Zu beachten ist aber, dass bei standardmässigem *TM* die Wortreihenfolge keine Rolle spielt: als einzige Grenzen gelten Dokumente. Die Wörter, die innerhalb der Dokumente auftreten, werden als bag-of-words aufgefasst, so dass zwei Dokumente, die dieselben Wörter in der gleichen Auftretenshäufigkeit haben, diese jedoch in zwei völlig unterschiedlichen Reihenfolgen führen, als identisch aufgefasst werden, selbst dann, wenn sie etwas (wesentlich) anderes bedeuten.[92]

Bei *LDA* können diverse Parameter eingestellt werden, wobei die wichtigsten Parameter diese zwei sind:

- Die Anzahl der gewünschten Topics.
- Die Anzahl Wörter, die genutzt werden, um einen Topic zu konstituieren.

Als gutes Beispiel, zu was *TM* mit dem *LDA*-Verfahren fähig ist, bietet sich ein Beispiel des "Content Mapping mit Topics Models" von Joachim Scharloth in seinem Blog [103] an, wo er die Themenschwerpunkte von "linken Szenen" anhand inhaltlicher Suchbegriffe einer indymedia.org-Webseite herausarbeitet und diese auch geografisch (für Deutschland) verortet. Einführend argumentiert er, dass beim *BND* (möglicherweise) eingesetzte Verfahren der inhaltlichen Suche direkt als Analyse gelten können, wenn sie mittels Verfahren des *TM* vollzogen werden.

3.5.2 Kombinationen zu 2, 3, 5, 8 und 10 Worten

Im Rahmen dieser Arbeit wurde, um eine unbegründete Willkür der Auswahl spezifischer Topics einzuschränken, auf Standardeinstellungen des eingesetzten *TM*-Werkzeugs[93] gesetzt und die Anzahl Topics zum einen fest auf 5 eingestellt, und die Zahl der Wörter

[90] Es existiert eine Erweiterung, die als probabilistic Latent Semantic Analysis *pLSA* bekannt ist.

[91] Bei *LSA* werden Dokumente einer Kollektion mitsamt ihren Termen in eine Matrix-Repräsentation überführt und mittels Verfahren der Matrizen-Dekomposition in kleinere Bestandteile aufgespalten – mit Matrizenmultiplikation wird der Ursprung wieder hergestellt. Durch diese Komplexitätsreduktion werden wichtige – latent zusammenhängende – Terme sichtbar. Diese konstituieren genauso wie bei *LDA* einen Bedeutungszusammenhang, deren Güte aber nicht optimal ist, wie Blei in seinem Paper darlegt.

[92] Vgl. kritisch hierzu Bubenhofer und Scharloth (2015). [11, S.13]

[93] Zum Einsatz kam das *Topic Modeling Tool*, das das bekannte *mallet*-Backend verwendet, *LDA* (mit gibbs-Sampling) zu betreiben. Vgl. URL `https://code.google.com/p/topic-modeling-tool/`. Abruf: 15. November 2015.

für jeden der Topics entsprechend auf 2, 3, 5, 8 oder 10 gesetzt. Das heisst, dass mit Wortkombinationen von mindestens zwei, höchstens aber zehn Worten gesucht wird.

Selektoren, die durch dieses Verfahren generiert werden, sehen semantisch vielversprechend[94] aus; je ein Beispiel:

- Ein 5-Wort-Topic aus dem *Aufbau*-Korpus: {`frauen`, `zürich`, `kapitalismus`, `märz`, `demo`}
- Ein 10-Wort-Topic aus dem *PNOS*-Korpus: {`schweizer`, `volk`, `schweiz`, `politik`, `volkes`, `leben`, `grund`, `armee`, `meinung`, `politische`}

3.6 Selektorenauswahl für die Methoden 1 und 2

Während bei der Überwachungsmethode 3 von Sektion 3.5 direkt nur fünf Selektoren für jedes der Selektorentypen von 2, 3, 5, 8 oder 10 Worten generiert wird, muss die Auswahl bei den Überwachungsmethoden 1 und 2 gemäss Sektionen 3.3 und 3.4 zunächst noch erfolgen.

Es wird dabei wie folgt vorgegangen:

1. Für jeden Ausdruck[95] t und Dokument d der Textkollektion D des einen oder anderen Trainingskorpus' wird der TFIDF-Wert berechnet.
2. Für jedes Dokument d werden die fünf höchsten TFIDF-Werte in einer eigenen Datei gespeichert.
3. Über alle Dateien aus 2. werden diejenigen fünf t-Ausdrucke ausgewählt, die absolut – über die gesamte Kollektion verteilt – am häufigsten vorkommen.

[94] Andere Selektoren sind unter A.1.3 zu finden.

[95] Das können entsprechend den Modellen Einzelworte, Wort-2-Gramme oder Wortkombinationen (bis zu fünf Worte) sein.

4 Evaluation der Selektoren

> *Wenn also diese Daten offensichtlich nicht ausreichen, um einen Anschlag zu verhindern – welche Daten um alles in der Welt hofft man dann per Generalüberwachung zu bekommen? Die rationale Herangehensweise wäre das Eingeständnis, dass es nicht darum geht, neue Daten zu bekommen, sondern die längst vorhandenen besser auszuwerten. Die scheinrationale Herangehensweise aber wird sich durchsetzen: mehr Überwachung. Mehr Daten. Die Irrationalität dahinter lautet: Wir finden die Nadel im Heuhaufen nicht, also brauchen wir mehr Heu.*
>
> – Sascha Lobo am 25. November 2015 [63]

Dieses Kapitel führt auf der einen Seite eine Simulation von *XKeyscore* (mit beschränkten Funktionsumfang) und öffentlich (web-)indexiertem Material durch, um aufzuzeigen, welche Art von Inhalten angesteuert werden müssten, um bei der Massenüberwachung in die engere Auswahl zu kommen; auf der anderen Seite wurde auch ein eigener Datenstrom für zehn Tage erzeugt, um ein realistisches Überwachungsszenario herbeizuführen, reale private Daten zu überwachen sowie Sparse-Data-Problematiken zu thematisieren.

4.1 Zum Umfang der Evaluation

Die Evaluation hat den folgenden Umfang:

- In der Summe der drei Modelle zur Massenüberwachung werden 70 Selektoren je Trainingskorpus (*Aufbau*; *PNOS*) evaluiert: in der Summe sind das 140 Selektoren.
- Es werden vier Suchmaschinen als Evaluationskorpora genutzt und der eigens erstellte Datenstrom als weiteres (persönlich differenziertes) Evaluationskorpus beigezogen: insgesamt fünf Evaluationskorpora.
- Im Produkt ergeben sich 700 zu vollziehende Selektionen.
  - Bei den Suchmaschinen werden je Selektion (oder Suchabfrage) auf der ersten Ergebnisseite die fünf obersten Ergebnisse ausgewählt, sofern sie (1) nicht direkt das Trainingsmaterial reflektieren oder (2) unplausibel sind. Ersteres bedeutet, dass ein durch das PNOS-Trainingsmaterial generierter Selektor dann einen ungültigen Treffer produziert hat, wenn dieser Treffer ein Treffer unter pnos.ch

darstellt oder die Inhalte offensichtlich direkt von da stammen.[96] Zweitens ist ein Treffer dann ungültig, wenn bei der Betrachtung des Treffers auffällt, dass auf Grund der dynamischen Natur[97] der Webseite oder wegen Eigenheiten der Suchmaschine, der Treffer unsinnig ist.

– Beim überwachten Datenstrom wird analog unter Einsatz einer lokalen Suchmaschine vorgegangen.

– Je nach Spezifizität der Selektoren und der Grösse und begrifflichen Breite der Evaluationskorpora sind damit – gegeben (gültige) Treffer – bis zu 3'500 Ergebnisse möglich, die zu evaluieren wären.

4.2 Die Evaluationssysteme und -daten im Einzelnen

Die vier Suchmaschinen und der eigene Datenkorpus – entstanden durch Überwachung der eigenen Internetleitung – werden hier soweit nötig bekannt gemacht. Unbeantwortet bleibt in einigen Fällen – wie bei *StartPage* oder *DuckDuckGo* – die Frage, wie gross das Datenkorpus jeweils ist.

4.2.1 Whitenet-Index: DuckDuckGo

Bei *DuckDuckGo* handelt es sich um eine Suchmaschine, die als alternative zur *Google*- oder *StartPage*-Suchmaschine insbesondere im Nachgang der Snowden-Enthüllungen populär geworden ist und gemäss eigenem Fact-Sheet weder die eigenen User überwacht noch eine "filter bubble" entstehen lässt, wonach Benutzer auf Grund ihres eigenen Surf- und Suchverhaltens oder anderer Merkmale, mit spezifischen Ergebnissen konfrontiert werden. [19]

Die Ergebnisse sollen sich aus 400 verschiedenen Datenquellen zusammensetzen, die allerdings nicht im Detail benannt werden, so dass keine Klarheit darüber besteht, wie gross der Datenkorpus – in Anzahl (indirekt) indexierter Dokumente – ist. [20]

4.2.2 Whitenet-Index: StartPage

Bei *StartPage* handelt es sich um eine Suchmaschine, die im Backend die weltweit populäre *Google*-Suchmaschine verwendet: sie hat gegenüber direkten *Google*-Suchen den Vorteil, dass die Suchabfragen frei von personenbezogenen Details (anonymisiert) abgesetzt werden.

[96] Ein Treffer ist allerdings valide, wenn ein Selektor zum Beispiel aus dem Aufbau-Korpus induziert ist, und einen pnos.ch-Treffer erzeugt.

[97] Damit ist gemeint, dass sich die Inhalte offensichtlich häufig ändern: beispielsweise bei Feeds oder bei Webseiten, die sich mit jedem Abruf ändern.

Insbesondere sollen weder Cookies noch IP-Adressen bei Suchabfragen gespeichert werden. [112]

Dies ist deshalb wichtig, weil die *Google*-Suchmaschine bekanntlich über viele Kriterien verfügt, Suchbegriffe entsprechenden -resultaten zuzuordnen. Browser-, Personen-, Länder-bezogene Effekte werden unter Einsatz von *StartPage* möglichst minimiert.

Die Grösse des Suchindexes ist nicht bekannt, weil dieser direkt von dem von *Google* abhängt. Allerdings eröffnet die Suche mit häufigen Einzelbuchstaben wie “a” oder “e” , dass es sich um mindestens Milliarden von indexierten Inhalten handeln muss.

Was beachtet werden muss, ist dass durch Regelungen wie das “Recht auf Vergessen” in der Europäischen Union oder eigenen Vorstellungen, angenommen werden muss, dass viele Inhalte, die eigentlich im Whitenet existieren, nicht angezeigt werden, weil sie zensiert sind.

4.2.3 Darknet-Index: Not Evil

Mit *Not Evil* besteht eine Suchmaschine im *Darknet* des *Tor*-Netzwerks, die per 30. November 2015 unter der URL `http://hss3uro2hsxfogfq.onion/` abrufbar ist. Eigenen Angaben – zuunterst auf der Suchmaschine gemäss – sind rund 15.8 Millionen Links indexiert, deren URLs auf die Endung “.onion” enden, weil *Not Evil* strikte nur nach Inhalten sucht, die als sogenannte Hidden-Services betrieben werden. Das sind Serverdienste mit beispielsweise Webinhalten, deren genauer Standort auf Grund des Onion-Routing-Prinzips des *Tor*-Netzwerks nicht ohne Weiteres bekannt ist.

4.2.4 P2P-Index: YaCy

YaCy ist eine Suchmaschine nach dem peer-to-peer-Prinzip, wonach die indexierten Inhalte nicht – wie bei den anderen Diensten – zentral von einem Anbieter kommen, sondern dezentral von verschiedenen peers zur Verfügung gestellt und laufend verändert werden.

Im offen zugänglichen Netzwerk “freeworld” sind bei letzter Betrachtung[98] – je nach Zeitraum – zwischen 313 (Tagesbasis) und 1’106 (Monatsbasis) peers daran beteiligt, Dokumente für eine Websuche zu indexieren. Die Anzahl indexierter Dokumente beträgt knapp 2 Milliarden.

Wird *YaCy*-Software auf dem eigenen Rechner installiert, so können standardmässig unter der URL `http://localhost:8090` direkt Websuchen abgesetzt werden. Zudem ist es möglich, sich selber als peer zu beteiligen und am dafür nötigen Web-Crawling teilzunehmen,

[98] Stand: 30. November 2015.

den Webindex zu vergrössern.

Im Gegensatz zu anderen Suchmaschinen kann bei *YaCy* – nicht zuletzt durch den Fakt, dass es sich um quelloffene Software handelt – davon ausgegangen werden, dass die Ergebnisse zensurfrei sind.

4.2.5 Privat-Index: Eigener 10-Tages-Datenstrom "Own"

Um ein persönlich differenziertes Evaluationskorpus (im Folgenden *Own* genannt) aufzubauen, das insbesondere einem realen Szenario entspricht, wie es zum Beispiel mit der Schweizer Kabelaufklärung nach Inkraftsetzung des *Nachrichtendienstgesetzes* möglich würde, wurde die Idee umgesetzt, eigenen Datenverkehr mitzuschneiden.[99]

Über 10 Tage[100] habe ich meinen eigenen[101] Internetanschluss in Selbstüberwachung unter Einsatz eines *alix*-Boards[102] mitgeschnitten. Zum Einsatz kam das Werkzeug *tcpflow(1)*[103], das Dekoder dafür bietet, aus den erfassten Rohdaten in Echtzeit Dokumente wie Bilder, PDF- oder HTML-Dateien zu extrahieren.

Insgesamt wurden 21GB Daten erfasst, verteilt auf 256'097 Dateien, ohne Duplikate deren 212'045.[104] 91'210 Dateien sind gemäss *file(1)*-Kommando [105] Daten rein binärer Natur, ohne bekanntes Dateiformat: bei der Betrachtung ohne Duplikate bleiben 89'873 Dateien.[106] Diese sind linguistisch nicht ohne Weiteres zugänglich und mit den den hier genutzten Selektoren nicht durchsuchbar[107], wenn auch trotzdem textuelle Inhalte in diesen Dateien enthalten sein können. Wird mittels *strings(1)*-Kommando das eigentlich zu analysierende Korpus[108] in eine rein textuell fassbare Version überführt, zählt der Korpus 208'577 Dateien.[109]

[99] Die mitgeschnittene Sicht entspricht der eines Internet-Zugangsproviders und aller weiteren Akteure, die dessen Datenströme erhalten.

[100] Vom 29. September bis zum 9. Oktober 2015.

[101] Von meiner Familie und mir – insbesondere meiner Lebenspartnerin und mir – genutzt.

[102] *AMD Geode*-Gerät mit 256MB RAM und *FreeBSD 5.2-RELEASE*-Betriebssystem.

[103] Vgl. *github*-Repository. URL `https://github.com/simsong/tcpflow`. Abruf: 23. November 2015.

[104] Allerdings bei etwa gleichbleibender Gesamtdateigrösse von rund 21GB.

[105] Vgl. im Anhang D.1.1

[106] Vgl. im Anhang D.1.2

[107] Zu betonen an dieser Stelle aber sei, dass beispielsweise *XKeyscore* fähig ist, mittels "fingerprints" Muster in binären Datenströmen zu erkennen, die gegebenfalls eine (weitergehende) Dekodierung erlauben, um linguistisch verwertbare Daten freizugeben.

[108] Ohne Duplikate

[109] Diese Form des *Own*-Korpus ist im Ordner `data/eval/fulltake/4_raw_strings_uniq` abgelegt: die Gesamtgrösse des Korpus nur mit Text beträgt rund 12% der ursprünglich rohen Grösse: 2.6GB. Einige Dateien werden zu Duplikaten, weil durch die rein textuelle Repräsentation ihre Substanz identisch wird. Das können Dateien sein, die sich binär unterscheiden, in ihren textuell repräsentierten (und spärlichen) Meta-Informationen hingegen identisch sind. Auch können Duplikate dadurch entstehen, dass unmittelbar keine textuellen Informationen extrahierbar sind.

Zur Indexierung und die Suche wurde die quelloffene, lokal installierbare Suchmaschine *Recoll*[110] genutzt, welche Treffer nach dem BM25-Algorithmus[111] sortiert.

4.3 Manuelle Annotation

4.3.1 Durchführung der manuellen Annotation

Der Annotation zu Grunde standen Treffer in Form von URLs in einer ODS-Tabelle. Die Annotatoren hatten die Aufgabe, jeden dieser Treffer als entweder "links-" oder "rechtsextremistischen" Inhalt einzustufen[112] und ihn damit als True-Positive zu qualifizieren.

Obwohl jede URL eindeutig einem bestimmten Überwachungsmodell und innerhalb davon einem konkreten Aufbau- oder PNOS-induziertem Selektor zugewiesen werden kann, wurde zur Verhinderung durch diese Informationen beeinflusster Entscheide, jeder Bezug zu solchen Angaben entfernt und die URL-Liste zudem alphabetisch geordnet. Damit sollte verhindert werden, dass (1) (stark unterschiedliche) URLs bloss auf Grund ihrer Nachbarschaft ähnlich annotiert werden und (2) umgekehrt Annotationen aber vereinfacht werden, die offensichtlich schon auf Basis der URL miteinander zusammenhängen: in Fällen, wo beispielsweise Unterseiten einer Webseite referenziert werden, war diese Präsentation für eine konsistente Annotation hilfreich.

Grundsätzlich existieren drei Arten von URLs:

- Gewöhnliche URLs, die offen zugänglich sind.
- URLs mit dem Host-Namen "alix", welche Ergebnisse des eigenen Evaluationskorpus *Own* darstellen.
- URLs mit der Domain-Endung ".onion", die nur über das *Tor*-Netzwerk zugänglich sind. Diese stehen im Zusammenhang mit Treffern aus dem *Not Evil*-Korpus.

Als Annotatoren haben der Autor (im Folgenden als Annotation H) und seine Lebenspartnerin (im Folgenden als Annotation S) gearbeitet: es wurde unabhängig und ohne Absprache annotiert.

Zu betonen ist allerdings, dass vorgängig für die Annotationen[113] H und S 770 Treffer

[110] Vgl. Webseite von *Recoll*. URL http://www.lesbonscomptes.com/recoll/. Abruf: 23. November 2015.

[111] Eine TFIDF-Variante, um Suchresultate zu sortieren.

[112] Dafür wurde in einem Feld "L" oder "R" (exklusiv) eine 1 eingetragen, um eine einfache Zählung zu erlauben.

[113] Ausgeführt allerdings nur vom Annotatoren H, womit nicht ausgeschlossen werden kann, dass einige True-Positives mehr möglich gewesen wären: dies (1) allerdings auf dafür sehr fragwürdigen Seiten beziehungsweise mit äusserst fragwürdiger Urheberschaft (für eine Einstufung als "links-" oder "rechtsextrem") und (2) mit sehr geringer Wahrscheinlichkeit.

oder rund 34.53% aller validen Treffer manuell als False-Positives markiert wurden, weil diese offenkundig nichts mit “Links-” oder “Rechtsextremismus” zu tun haben sollten[114], sondern allenfalls darüber berichten; es sind dies Inhalte (vor allem HTML-Dateien) der folgenden und abschliessenden Art:

- Konkrete Blogs, die beiden Annotatoren bekannt sind und entweder dem CCC- oder der Piratenpartei Schweiz zuzuordenen sind: blog.fdik.org, blog.fefe.de, blog.fukami.io und substanz.davidherzog.ch.
- Verbreitete, bekannte Zeitungen[115] , bis mithin “Systemzeitungen”[116]: 20min.ch, abendblatt.de, ardmediathek.de, *.ard.de, arte.tv, bazonline.ch, bernerzeitung.ch, blick.ch, br.de, bzbasel.ch, derbund.ch, derstandard.at, diepresse.com, dw.com, golem.de, focus.de, handelsblatt.com, handelszeitung.ch, kleinezeitung.at, krone.at, futurezone.at, hochschulanzeiger.faz.net, tagesanzeiger.ch, tageswoche.ch, rbb-online.de, rp-online.de, schweizer-illustrierte.ch, *.sonntagszeitung.ch, spiegel.de, srf.ch, stern.de, stuttgarter-zeitung.de, *.sueddeutsche.de, watson.ch, *.nzz.ch[117], *.berliner-zeitung.de, vice.com, zol.ch und swissinfo.ch.
- Metaseiten, die News indexieren, sofern diese als solche offensichtlich erkannt wurden: news.google.*, pressekompass.net sowie Einzelseiten im *Own*-Datenkorpus.
- Regierungsseiten oder Staatsplattformen: baselland.ch, *.be.ch, ch.ch, bern.ch, bpb.de[118], *.admin.ch, erlangen.de, parlament.ch, *.bundestag.de, stadt-zuerich.ch, *.bund.de, *.nrw.de, vaud.ch, vd.ch, *.bs.ch, glarus.ch und deggendorf.de.
- Gerichtsseiten: bundesverfassungsgericht.de
- Firmenwebseiten: schweizer-metallbau.ch, toppreise.ch, zugergrafik.ch, zattoo.com, sudwerk.ch, microspot.ch, naturkundemuseum-berlin.de und home.ch.
- Wohlbekannte Community-Webseiten: debian.org, holarse-linuxgaming.de, *.wikipedia.org, php.net, selfphp.net und *.freifunk.net.
- Verbände: Amnesty[119], FSFE[120], konsumentenschutz.ch, humanrights.ch, redcross.ch

[114]Es kann dies freilich nicht völlig ausgeschlossen werden.

[115]Wenn offenkundig Forenbereiche oder Kommentarbereiche von Online-Artikel referenziert werden, sind diese in der Liste nicht enthalten: sie wurden der doppelt durchgeführten Annotation auf Einzelbasis unterworfen. Es ist allerdings zu beachten, dass beispielsweise gemäss Aussagen im *SRF Medienclub* vom 10. November 2015 auf Basis von unbekannten Stichworten beträchtliche Anteile von Kommentaren bei Online-Zeitungen ausgesondert werden, wo vermutet wird, dass die Aussagen einer unbekannten Norm abweichen. Konkret wird das in einem Blog-Eintrag von Volker Birk aufgezeigt. [5]

[116]Damit sind Zeitungen gemeint, die vorwiegend die herrschende Meinung repräsentieren und Minderheitsmeinungen wenig, keine oder nur negative Publizität bieten. Es kann angenommen werden, dass sie mithin die Funktion erfüllen, das herrschende System zu stützen und damit diametral den Interessen von “Links-” und “Rechtsextremisten” entgegenstehen.

[117]Darunter auch beispielsweise campus.nzz.ch.

[118]Bundesamt für politische Bildung [Deutschland].

[119]In Form von zwei YouTube-Videos.

[120]Eine HTML-Seite im *Own*-Korpus.

und transparency.ch.

- Diese Bücher[121] sind auch direkt als False-Positives markiert: “Common Wealth. Das Ende des Eigentums” (Hardt et al. 2010), “Privacy-Handbuch” (Doctorow et al. 2013), “Little Brother” (Doctorow 2009), “Inside Wikileaks” (Domscheit-Berg 2011), “Das Neue Spiel. Strategien für die Welt nach dem digitalen Kontrollverlust” (Seemann 2014), “Die Satanischen Verse” (Rushdie 1989 [1988]), “Ulysses” (Joyce 1975)
- UZH-Forschungsdokumente vom RWI[122], dem IPZ[123], dem IVR[124] und dem Psychologischen Institut.
- Forschungspublikationen: Einerseits ein Soziologie-,[125] andererseits ein Linguistik-Paper.[126]
- Weitere Dokumente universitärer Hochschulen: Bamberg, Münster, Siegen, Köln, Konstanz, Stuttgart, Leipzig, Hamburg und Wien.
- Publikationen von Verlagen: link.springer.com
- Weitere wissenschaftliche Publikationen: arxiv.org
- Seiten für touristische Zwecke oder Reisebuchungen: booking.com, swisscommunity.org, myswitzerland.com, fluege.de und flughafen-zuerich.ch.
- Wörterbuch- oder Übersetzungsseiten: duden.de, *.wiktionary.org, dictionary.reference.com, dictionary.reverso.net, *.thefreedictionary.com, dict.leo.org, linguee.de und wirtschaftslexikon.gabler.de.
- Webseiten von (Schweizer) demokratischen Volksinitiativen, wo Volksrechte im Rahmen der Verfassung wahrgenommen werden: Konzernverantwortungsinitiative und Wohn-Initiative.
- Nicht-linguistisches Textmaterial, JavaScript-Code und -Kommentare, CSS-Code und -Kommentare, blosse Wortlisten[127] oder die Meta-Informationen von Bildern.

Im Gegenzug wurden einige harmlos anmutende Webseiten oder Plattformen zur näheren Evaluation freigegeben, weil konkrete Beiträge “extremistisch” eingestuft werden kön-

[121] Sie sind allesamt dem *Not Evil*-Korpus zuzurechnen.

[122] Rechtswissenschaftliches Institut

[123] Institut für Politikwissenschaft

[124] Institut für Völkerrecht und ausländisches Verfassungsrecht

[125] Ironischerweise ein vom Autor dieser Arbeit selber geschriebenes Dokument, das sich mit *Liquid Democracy* beschäftigt und im Own-Korpus als PDF liegt: Liquid Democracy – Zwischen repräsentativer und direkter Demokratie. Grundlagen, Umsetzungen und Diskussion bezüglich Chancen und Gefahren. *Sociology in Switzerland. Democracy in Politics and Social Life.* August 2011. URL `http://socio.ch/demo/t_marques.pdf`. Abruf: 26. November 2015.

[126] Ausgerechnet ist dies die Publikation, die der Überwachungsmethode 2 dieser Arbeit als theoretische Grundlage dient. [21]

[127] Es wurden Dokumente unter cd.textfiles.com als Treffer erzeugt, die diesen Charakter aufweisen.

nen, diese konkret personenbezogene Webseiten von "extremistischen" Personen in der Einordnung der Geheimdienste enthalten könnten, wo sich diese selber darstellen oder aber Gruppen sich exponieren, die diese Plattform zur Präsentation, Kommunikation und Rekrutierung nutzen; im Fokus des Verdachts steht also der da vorhandene User-Generated-Content *UGC* und nicht die Plattform selber; solche Plattformen können sein:

- Social-Media-Plattformen: xing.com, facebook.com, youtube.com, vimentis.ch, reddit.com und bandcamp.com.
- Firmen: amazon.com[128]

Grenzfälle von Seiten, die zwar anerkannt und bekannt sind, dennoch (im Einzelfall Plattform für) "links-" oder "rechtsextremistische" Positionen enthalten können und deshalb zur näheren Evaluation freigegeben sind, bestehen ebenso und sind:

- Zeitungen: woz.ch, taz.de
- Parteien: pda.ch, svp.ch

4.3.2 Statistik der manuellen Annotation

Mit den 770 vorbestimmten False-Positives gemäss dem Abschnitt 4.3.1 gilt es zusätzlich 1'460 Inhaltselemente auf Einzelbasis seitens der zwei Annotatoren manuell zu überprüfen, wobei die Summen der Ja-Annotationen in Tabelle 4.1 festgehalten sind: H_L bezeichnet dabei die Anzahl der Ja-Annotationen für die Annotation H innerhalb vom "Linksextremismus", während S_R entsprechend die Anzahl Ja-Treffer bezeichnet, welche die Annotatorin S für den Bereich "Rechtsextremismus" gesehen hat.

H_L	S_L	H_R	S_R
379	350	103	62

Table 4.1: Manuelle Ja-Annotationen H und S nach "Links-" (L) und "Rechtsextremismus" (R)

Um die Übereinstimmung zu messen, wurde das Inter-Annotator-Agreement *IAA* auf Basis des Kappa-Koeffizienten[129] zwischen den zwei Annotationen H und S bestimmt und jeweils gesondert nach "links-" und "rechtsextremen" Treffern durchgeführt.

Die Formel hierfür lautet:

$$\kappa = \frac{p_o - p_c}{1 - p_c} \tag{4.1}$$

[128] Auf Grund der da vorhandenen Reviews und auf Grund der Tatsache, dass beispielsweise in den USA Bücher auf dem Markt sind, welche nach europäischem Verständnis dermassen "extremistisch" sind, dass sie keinen Verleger finden, in den USA aber auf Grund des dort stark gewichteten Grundsatzes des "free speech" durchaus publiziert werden mögen.

[129] Nach Cohen Kappa

p_o bezeichnet dabei die Wahrscheinlichkeit der konkreten Übereinstimmung zwischen H und S, wohingegen p_c die Wahrscheinlichkeit bezeichnet, dass H und S zufällig übereinstimmen.

Es resultieren die Erwartungswerte von Tabelle 4.2:

p_{0_L}	p_{0_R}	p_{c_L}	p_{c_R}
94.66%	96.46%	72.64%	92.86%

Table 4.2: Erwartungswerte H und S für “Links-” (L) und “Rechtsextremismus” (R)

Es wird deutlich, dass die Übereinstimmung überzufällig ist (insbesondere im “linksextremen” Bereich). Die Wahrscheinlichkeitsbereiche sind deshalb hoch, weil die meisten Inhalte auf Grund der relativ hohen Rate an False-Positives weder dem “Links-” noch dem “Rechtsextremismus” zufallen.

Die κ-Werte, die resultieren, sind in Tabelle 4.3 zu entnehmen, welche im “links-” eine substanzielle und im “rechtsextremen Bereich” eine zumindest moderate Übereinstimmung begründen. [60]

	κ_L	κ_R
Wert	0.8	0.5
Beurteilung	Substanzielle Übereinstimmung	Moderate Übereinstimmung

Table 4.3: Kappa-Werte “Links-” (L) und “Rechtsextremismus” (R)

4.4 Ergebnisse

In dieser Sektion werden die Ergebnisse der Evaluation, die manuell mittels zwei Annotationen durchgeführt wurde, quantitativ (statistisch) präsentiert. Die kritische Diskussion und Interpretation der Ergebnisse und auch deren Grundlagen erfolgt abschliessend in Kapitel 5.

4.4.1 Trefferstatistik

Von den maximal möglichen 3'500 zu beachtenden Treffer wurden 2'242 über alle Selektoren hinweg valide erzeugt: “bloss” 64.06% der maximal möglichen Treffer sind zur Evaluation somit überhaupt vorhanden, was heisst, dass in vielen Fällen ein Sparse-Data-Problem angezeigt ist, wobei die Unterschiede je nach Betrachungsweise (Evaluationssystem oder Überwachungsmodell) beträchlich sind.

4.4.1.1 Nach Evaluationssystem

Die Rangliste der meisten validen Treffer von den 700 je Evaluationssytem möglichen, wird – wenig überraschend – von *StartPage* angeführt, mit dem *Own*-Datenkorpus als Schlusslicht, wo es am wenigsten "Substanz" gibt, fündig zu werden, wie aus Tabelle 4.4 sichtbar wird.

Evaluationssystem	Treffer absolut (maximal)	Treffer relativ
StartPage	649 (700)	92.71%
DuckDuckGo	550 (700)	71.43%
YaCy	480 (700)	68.57%
Not Evil	339 (700)	48.43%
Own	224 (700)	32%
Total	2′242 (3′500)	64.06%

Table 4.4: Trefferstatistik: Nach Evaluationssystem

4.4.1.2 Nach Überwachungsmodell

Wird nach dem Überwachungsmodell und die dadurch erzielten Treffer gefragt, so rangiert gemäss Tabelle 4.5 zuoberst das TFIDF-Modell.

Überwachungsmodell	Treffer absolut (maximal)	Treffer relativ
TFIDF	593 (750)	79.01%
Verdachtssprache	992 (1′500)	66.13%
LDA	657 (1′250)	52.56%
Total	2′242 (3′500)	64.06%

Table 4.5: Trefferstatistik: Nach Evaluationssystem

Das Ranking muss allerdings aus verschiedenen Gründen mit Vorsicht betrachtet werden, weil mögliche Interpretationen der Güte deshalb nicht möglich sind, weil die Selektoren verschiedener Länge und auch Anzahl sind. Es bestehen weder alle Selektoren aus der gleichen Anzahl Wörter noch sind sie anderweitig in ihrer Natur direkt vergleichbar, noch kommen bei allen Modellen die gleiche Anzahl Selektoren zum Einsatz. Während die Anzahl der eingesetzten Selektoren – sofern sie sich in ihrem Wesen nicht fundamental unterscheiden – einen nur geringen Einfluss darauf haben mag, wie die Ergebnisse im Gesamtschnitt für das jeweilige Modell ausfallen, so ist die Tatsache, dass die Selektoren etwa des TFIDF-Modells wesentlich einfacher aufgebaut sind, sicherlich entscheidend dafür, zu erklären, weshalb diese mehr Ausbeute erreichen. Schliesslich gibt es beim TFIDF-Modell im einfachsten Fall Selektoren die aus simplen Einzelworten bestehen, wofür im Allgemeinen die Chance hoch ist, dass diese in einem der Evaluationskorpora anzufinden sind.

4.4.2 True-Positive-Statistik

Für die True-Positve-Statistik gilt es zu beachten, dass im Wesentlichen zwei Masse P (für Precision) und S (für Score) zum Einsatz kommen, um die Performance der Modelle und ihrer Selektoren zu messen:

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4.2}$$

$$S = \frac{P}{\frac{\text{VH}}{\text{PH}}} \tag{4.3}$$

Der Score-Wert S ist vor allem in denjenigen Fällen interessant, wo ein hoher P-Wert zwar erzeugt wird, dies aber unter der Bedingung der Fall ist, dass auch nur wenige Treffer überhaupt erzeugt wurden. Dies kann in Fällen, wo ein Korpus ein Sparse-Data-Problem hat, die Ergebnisse entzerren helfen.

Dabei gilt es ferner folgende Definitionen zu beachten, um die Tabellen im Folgenden interpretieren zu können:

- TP: Anzahl True-Positives
- FP: Anzahl False-Positives
- VH: Anzahl Valid-Hits oder gültiger Treffer (von den maximal möglichen Hits PH)
- PH: Anzahl Possible-Hits oder wieviele Treffer maximal mit einem Selektor hätten gültig erzeugt werden können
- P_h: Precision gemäss manueller Bewertung des Annotators H
- P_s: Precision gemäss manueller Bewertung der Annotatorin S
- $P_{h \cap s}$: Precision entsprechend der gemeinsamen TP- und FP-Übereinstimmung der Annotationen H und S oder die eigentlich untere Precision-Schranke
- S_h: Score gemäss manueller Bewertung des Annotators H auf Basis der bekannten eigenen Precision P_h
- S_s: Score gemäss manueller Bewertung der Annotatorin S auf Basis der bekannten eigenen Precision P_s
- $S_{h \cap s}$: Score entsprechend der gemeinsamen TP- und FP-Übereinstimmung der Annotationen H und S auf Basis der bekannten Precision $P_{h \cap s}$ oder die eigentlich untere Score-Schranke
- $\overline{P}$: Precision im Durchschnitt aller anderen Precision-Werte P_h, P_s und $P_{h \cap s}$, womit mögliche Fehlentscheidungen in den Annotationen H und S relativert werden können
- $\overline{S}$: Score im Durchschnitt aller anderen Score-Werte S_h, S_s und $S_{h \cap s}$, womit mögliche Fehlentscheidungen in den Annotationen H und S relativiert werden können

- $\overline{PS}$: Durchschnittswert der Precision-Werte $\overline{P}$ und $\overline{S}$, der im Idealfall die Gesamtergebnisse reflektiert und für das Ranking der Tabellen genutzt wird

4.4.2.1 Nach Evaluationssystem

Nach Evaluationssystem gemäss Tabelle 4.6 fallen zwei Ergebnisse deutlich auf: für das *Own*-Korpus konnten praktisch[130] keine True-Positives gefunden werden. Hingegen ist die *StartPage*-Kollektion ergiebiger: mindestens 20% der vorhandenen und validen Treffer VH sind True-Positives.

Eine erstaunlich hohe untere Schranke für die Precision bietet zudem das *Not Evil*-Korpus, wobei diese allerdings (korrigiert nach dem Verhältnis VH–PH) und ausgedrückt in Score-Werten relativiert werden muss. Zwar waren mindestens 23% der Treffer True-Positives, doch waren – wegen einem Sparse-Data-Problem – auch weniger Daten vorhanden, verdächtigen Inhalt zu finden.

Evaluationssystem	P_h	P_s	$P_{h\cap s}$	S_h	S_s	$S_{h\cap s}$	$\overline{P}$	$\overline{S}$	$\overline{PS}$
StartPage	0.25	0.21	0.20	0.22	0.19	0.19	0.22	0.20	0.21
Not Evil	0.25	0.24	0.23	0.11	0.10	0.10	0.24	0.10	0.17
DuckDuckGo	0.24	0.11	0.09	0.14	0.09	0.08	0.15	0.10	0.13
YaCy	0.16	0.11	0.09	0.10	0.08	0.06	0.12	0.08	0.10
Own	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4.6: True-Positive-Statistik: Precision- und Score-Werte nach Evaluationsdaten

4.4.2.2 Nach Überwachungsmethode

Die globale Betrachtung nach den drei genutzten Überwachungsmethoden zeigen gesamthaft in Tabelle 4.7 auf, dass die *LDA*-Modelle im Durchschnitt ihrer konkreten Selektoren am besten funktionieren, True-Positives zu generieren. Zu beachten ist aber, dass – rein qualitativ – die Güte einzelner der Selektoren, welche den anderen zwei Modellen zugefallen sind, fragwürdig sind. Eine andere Modellierung der Überwachungsmethoden könnte hier also andere Ergebnisse erzeugen.

Evaluationssystem	P_h	P_s	$P_{h\cap s}$	S_h	S_s	$S_{h\cap s}$	$\overline{P}$	$\overline{S}$	$\overline{PS}$
LDA	0.26	0.16	0.15	0.16	0.12	0.11	0.19	0.13	0.16
TFIDF	0.15	0.14	0.13	0.10	0.09	0.08	0.14	0.09	0.12
Verdachtssprache	0.12	0.11	0.09	0.08	0.07	0.06	0.11	0.07	0.09

Table 4.7: True-Positive-Statistik: Precision- und Score-Werte nach Überwachungsmodell

[130] Es besteht die Ausnahme zweier Treffer seitens der Annotatorin S, wobei diese im Gesamtergebnis (nach zwei Kommastellen gerundet) wegfallen.

4.4.2.3 Precision im Durchschnitt: Top-10

Wird der Fokus wie in Tabelle 4.8 darauf gelegt, nach dem höchsten Durchschnittswert der Precision $\overline{P}$ zu fragen, ist festzustellen, dass nicht ein *LDA*-Modell zuoberst gerankt ist, sondern solche der *Verdachtssprache* und der *TFIDF*-Überwachungsmethodik: dies allerdings bezogen auf das *Not Evil*-Korpus, das bekanntlich ein Sparse-Data-Problem haben könnte.

$\overline{P}$	**Überwachungsmodell**	**Evaluationssystem**
0.70	Extrem-Intensivierer (Verdachtssprache)	*Not Evil*
0.63	Wort-2-Gramme (TFIDF)	*Not Evil*
0.59	LDA-10 (LDA)	*StartPage*
0.58	LDA-8 (LDA)	*StartPage*
0.43	Hoch-Intensivierer (Verdachtssprache)	*Not Evil*
0.41	Wort-Kombinationen (TFIDF)	*Not Evil*
0.39	LDA-5 (LDA)	*StartPage*
0.37	LDA-5 (LDA)	*DuckDuckGo*
0.36	Wort-2-Gramme (TFIDF)	*YaCy*
0.35	LDA-2 (LDA)	*YaCy*

Table 4.8: Precision im Durchschnitt: Top-10

4.4.2.4 Score im Durchschnitt: Top-10

Die besten Scores im Durchschnitt ($\overline{S}$), die über alle 70 verschiedenen Varianten von Überwachungsmethode, entsprechender Modellierung und Evaluationskorpus vorhanden sind, machen solche aus, die mit dem *StartPage*-Evaluationskorpus im Zusammenhang stehen: bei diesem kann damit gerechnet werden, dass Milliarden von Dokumenten vorhanden sind. Entsprechend beeindruckt auch, dass jene Modellierungen der *LDA*-Überwachungsmethode die besten Ergebnisse erzeugen, die mit mehr – thematisch zusammenhängenden – Worten auskommen. In den ersten vier Plätzen befinden sich die drei *LDA*-Modellierungen, die mit fünf, acht und zehn Worten auskommen. Die Wort-Kombinationen nach *TFIDF*-Überwachungsmethodik, die den dritten Platz einnehmen, "profitieren" vermutlich von einer günstigen Wortwahl im Sparse-Data-Korpus von *Not Evil*.

$\overline{S}$	Überwachungsmodell	Evaluationssystem
0.47	LDA-10 (LDA)	*StartPage*
0.46	LDA-8 (LDA)	*StartPage*
0.37	Wort-Kombinationen (TFIDF)	*Not Evil*
0.33	LDA-5 (LDA)	*StartPage*
0.28	LDA-2 (LDA)	*StartPage*
0.26	LDA-3 (LDA)	*DuckDuckGo*
0.25	LDA-2 (LDA)	*YaCy*
0.25	LDA-3 (LDA)	*StartPage*
0.25	Skandalisierer (Verdachtssprache)	*StartPage*
0.24	LDA-2 (LDA)	*DuckDuckGo*

Table 4.9: Score im Durchschnitt: Top-10

4.4.2.5 Precision+Score im Durchschnitt: Top-10

In Tabelle 4.10 werden die Ergebnisse nach dem $\overline{PS}$-Mass betrachtet, was dem Durchschnitt aller Einzelmasse entspricht. Diese "globale" Sicht bietet einen interessanten Abschluss, denn werden die ersten fünf Ranking-Ergebnisse begutachtet, fällt auf, dass alle Überwachungsmethoden *LDA*, *Verdachtssprache* und *TFIDF* vertreten sind, wobei – wie zu erwarten – die thematischen *LDA*-Modelle am besten abschneiden. Bei den anderen Modellen bleibt die Frage offen, ob diese in ihrer ausgewiesenen Güte nicht zufällig davon "profitieren", sich am *Not Evil*-Korpus behaupten zu können, das einige Besonderheiten hinsichtlich der indexierten Webseiten, doch auch seines Umfangs aufweist. Denn: im weiteren Ranking ist nur noch einmal ein Nicht-*LDA*-Modell an 9. Stelle vorhanden, das aber wiederum auf Basis des *Not Evil*-Korpus "besteht".

$\overline{PS}$	Überwachungsmodell	Evaluationssystem
0.53	LDA-10 (LDA)	*StartPage*
0.52	LDA-8 (LDA)	*StartPage*
0.42	Extrem-Intensivierer (Verdachtssprache)	*Not Evil*
0.39	Wort-Kombinationen (TFIDF)	*Not Evil*
0.36	Wort-2-Gramme (TFIDF)	*Not Evil*
0.36	LDA-5 (LDA)	*StartPage*
0.30	LDA-2 (LDA)	*StartPage*
0.29	LDA-5 (LDA)	*DuckDuckGo*
0.29	Hoch-Intensivierer (Verdachtssprache)	*Not Evil*
0.29	LDA-2 (LDA)	*StartPage*

Table 4.10: Precision+Score im Durchschnitt: Top-10

5 Schlussbetrachtungen

Jeder chinesische Bürger soll laut aktuellen Berichten ein Punktekonto ("Citizen Score") bekommen, das darüber entscheiden soll, zu welchen Konditionen er einen Kredit bekommt und ob er einen bestimmten Beruf ausüben oder nach Europa reisen darf. In diese Gesinnungsüberwachung ginge zudem das Surfverhalten des Einzelnen im Internet ein – und das der sozialen Kontakte, die man unterhält.

[...]

[Es] [...] wird immer deutlicher, dass wir alle [auch im Westen] im Fokus institutioneller Überwachung stehen, wie etwa das 2015 bekannt gewordene "Karma Police"-Programm des britischen Geheimdienstes zur flächendeckenden Durchleuchtung von Internetnutzern demonstriert.

– Dirk Helbling et al. im "Das Digital Manifest" vom 12. November 2015 [39]

5.1 Diskussion des Versuchs und der Ergebnisse

5.1.1 Zur Güte der Evaluationsergebnisse

Die 140 Selektoren, die im Anhang unter A.1 aufgeführt sind, weisen weder quantitativ noch qualitativ denselben Charakter auf.

Wie die quantitativen Ergebnisse aus dem Kapitel 4 vermuten lassen, schneiden oft Selektoren aus den *LDA*-Modellen am besten ab, was heisst, dass sie verhältnismässig am meisten True-Positives generieren: das soll aber nicht darüber hinwegtäuschen, dass es nicht strukturell, sondern nur in Einzelfällen gelingt, die eingangs aufgeworfenen Hypothesen zu falsifizieren. Precision-Werte über 0.5 sind zwar vorhanden, allerdings nur für eine Handvoll der über 70 Möglichkeiten, die das Überwachungssetting bietet. Strukturell liegt die Rate der False-Positives bei über 50%, womit die Hypothesen halten. Insbesondere wird auch die dritte Hypothese dadurch erfüllt, dass es insbesondere unter Einsatz von spezifisch aufgebauten Selektoren, die wie bei den *LDA*-Modellen bis zu zehn Worte umfassen können, gelingt, bessere Ergebnisse zu erzielen – dies im Kontrast zu Selektoren, die nur aus einzelnen oder nur allgemeinsprachlichen, unspezifischen Worten zusammengesetzt sind.

Allerdings haben die verschiedenen Selektoren auch unterschiedliche Tendenzen und Güten in ihrer eigenen, qualitativen Natur: so scheint es mit den *TFIDF*-Modellen, die in den Gesamtergebnissen selten gut abschneiden, verschiedentlich zu gelingen, Named Entities zu erzeugen, wie folgende zwei Selektoren aus dem *Aufbau*-Korpus zeigen:[131]

- `ibrahim abdallah`: Diese eigentliche Named Entity (NE) nimmt Bezug auf Georges Ibrahim Abdallah, ein marxistisch gesinnter Aktivist, der für die Koordination diverser tödlicher Anschläge verurteilt wurde und bis heute in Frankreich (widerrechtlich) gefangen gehalten wird. [23]
- `marco camenisch`: Als NE referenziert dieser Ausdruck den bekannten Schweizer grün-anarchistischen Aktivisten, den die Basler Zeitung als *"Ikone der Zürcher Linksextremisten"* bezeichnet und dem die Haftenetlassung (wiederholt) erschwert wurde. [109, 123]

5.1.2 Willkürpotenzial in der Auswahl des Trainingsmaterials

Es muss kritisch betont werden, dass die konkrete Ausgestaltung der inhaltlichen Suchbegriffe, die als Selektoren generiert werden, im höchsten Masse von der Aufbereitungsqualität und Auswahl des Trainingsmaterials abhängt.

Will die politische Führung beispielsweise gewisse Gruppen auf Grund soziolinguistischer Merkmale wie Sprachgemeinschaft, -verhalten oder -schicht von der Massenüberwachung "verschonen", so kann sie das zu steuern versuchen, indem gewisse Daten aus dem Trainingskorpus gelöscht werden: damit wird zumindest verhindert, dass sich ein potenziell spezifischer Sprachgebrauch als Selektor manifestieren kann. [62, S.350ff.]

Zudem ist zu beachten, dass einige der konkreten Selektoren plausibel darauf hindeuten, dass personen-, gruppen-, orts- und aktualitätsspezifische Merkmale darin kodiert sind: so fallen in den Selektoren unter A.1 gelegentlich Begriffe wie `schweizer`, `rütli`, `pnos`, `flüchtlingsflut` und `obama`.

5.1.3 Willkürpotenzial in der Generierung der Selektoren

Im Rahmen dieser Arbeit wurden Überwachungsmethoden aus den Bereichen Information Retrieval, Korpuslinguistik und Topic Analysis ausgewählt, um Selektoren zu generieren.

Es können einerseits andere Methoden der automatischen Generierung von Selektoren eingesetzt werden, andererseits aber auch die bestehenden Modelle verändert werden – alles dies resultiert in anderen Selektoren, die ganz andere True-Positives und auch False-

[131] Parallel dazu werden auch im *PNOS*-Korpus und innerhalb des *TFIDF*-Überwachungsmethode Named Entities generiert

Positives erzeugen mögen.

In keinem Fall aber kann es gelingen, im Sinne eines "sprachlichen Fingerabdrucks" die perfekten Selektoren zu finden, welche nur True-Positives erzeugen. Denn (1) wird dies Menschen und auch Gruppen nicht gerecht, die sich weiterentwickeln und ihren Sprachgebrauch verändern; (2) kann kein Selektor dynamische Faktoren – wie zukünftige Tagesaktualitäten – in sich ohne Weiteres enthalten und (3) wurde im Rahmen der Selektorgenerierung in Kapitel 3 aufgezeigt, dass aus jedem Modell tausende von Selektoren möglich gewesen wären, um (Teile) der repräsentierten Textkollektion abzubilden. Es ist also ohnehin eine gewisse "Unschärfe" nötig, um andere gleiche oder ähnliche Dokumente in Datenströmen zu finden, die den festgelegten Suchkrtierien eines Auftrags der Massenüberwachung entsprechen: dies aber führt unweigerlich zu grossen Mengen von False-Positives. [104]

5.1.4 Offener Spielraum bei der Interpretation der Treffer

Bei den erfolgten Treffern – selbst wenn sich darunter True-Positives befinden – kann (grosse) Uneinigkeit darüber bestehen, ob diese der definierten Suchkategorie entsprechen: im Rahmen der Evaluation von Kapitel 4 ist zu Tage getreten, dass für Treffer des "Linksextremismus" zwischen den Annotatoren susbstanzielle Einigkeit darüber bestand, dass die Treffer der Suchkategorie entsprechen: nicht so aber für den Bereich des "Rechtsextremismus", wo der Kappa-Koeffizient mit 0.5[132] bedeutend tiefer ausgefallen ist.

Wird in Betracht gezogen, dass es noch diffusere Suchkategorien – wie etwa "Terrorismus" – gibt, die mithin stark ideologisch und – bei Kriegsparteien – standpunktspezifisch sein können, ist auch auf den "Faktor Mensch" zuletzt kein Verlass, dass keine Personen als Folge von Massenüberwachung Opfer weitergehender, invasiverer Formen der Überwachung werden, die ihr Leben (stark) negativ beeinflussen können.

5.1.5 Verhältnismässigkeit und Massenüberwachung

Der Umstand, dass Massenüberwachung eine ungezielte Form der Überwachung darstellt und auf den Grundsatz des Generalverdachts beruht, hat zur Folge, dass in realen Situationen Millionen Menschen verdächtigt werden und viele davon fälschlich ins Raster geraten, verdächtig zu sein. [6, 10]

Es sei entsprechend den Ergebnissen von Kapitel 4 darauf aufmerksam gemacht, dass die meisten Treffer harmlos sind: zu oft sind Newsseiten oder Wikimedia-Plattformen betroffen, womit beispielsweise journalistische Recherchearbeiten Verdacht schöpfen können.

Im Rahmen des Vorhabens *Mastering The Internet* des britischen Geheimdienstes *GCHQ*

[132] Gegenüber 0.8 für die Übereinstimmung beim "Linksextremismus"

fällt auf, dass insbesondere Termini wie "nationale Sicherheit" und "Wirtschaftsinteressen" sehr schwammig bis nicht definiert sind und entsprechend je nach Regierung und Tagesordnung anders ausgelegt werden können: daraus schöpft sich ihre Gefahr, die in massiv ausufernder Überwachung resultieren kann. [64]

Die jüngsten Anschläge von Paris haben – in teilweise paralleler Tonalität (von Vergeltung und Kriegsrhetorik) wie nach den Anschlägen von 9/11 – zu Forderungen nach mehr Überwachung geführt. Es gibt allerdings nach wie vor keinen plausiblen Hinweis darauf, dass ungerichtete Formen der Überwachung, die grundsätzlich darauf abzielt, alle zu verdächtigen – wie dies von der NSA und artverwandten Diensten betrieben wird – einen Terroranschlag im Sinne führen zu können, bislang zur Verhinderung derselben führen konnte. Dies bestätigen US-Geheimdienste in eigenen (internen) Papieren bisweilen selber. [41, 42]

Zudem wurde bei der höchstgerichtlichen Ausserkraftsetzung des sogenannten Safe-Harbor-Abkommens zur Datenübermittlung von Daten der EU[133] an die USA festgestellt, dass die durch die Snowden-Enthüllungen offengelegte Praxis der US-Massenüberwachung den *"[...] Wesensgehalt des [...] garantierten Grundrechts auf Achtung des Privatlebens [verletzt]"* und damit Grund darstellt, unwillentliche Datenübermittlungen in die USA zu stoppen. [108]

5.2 Strategien im Umgang mit der Massenüberwachung

Es ist einerseits theoretisch immer möglich, andererseits praktisch in Einzelfällen beobachtbar, Sprach- oder Verschlüsselungstechnologie einzusetzen, um die eigene Identität oder die Inhalte zu schützen. In Ländern, wo freiheitliche und für demokratische Prozesse erforderliche Güter wie Meinungsäusserungs- oder Versammlungsfreiheit nicht gewahrt sind und mit Repression seitens staatlicher Behörden gerechnet werden muss, können solche Werkzeuge direkt für das Überleben notwendig sein.

Zur Autorenidentifikation ist zudem eine Abgrenzung vorzunehmen, weil die Massenüberwachung nicht primär darauf abzielt, nach einer bestimmten Person zu suchen, sondern nach verdächtigen Personen(-Kreisen) generell beziehungsweise überhaupt interessierendem Material – anfänglich unabhängig der Autorenschaft.

[133] Die Schweiz hat ein eigenes, doch wesensgleiches Abkommen mit den USA abgeschlossen, das allerdings (noch) nicht ausser Kraft gesetzt wurde, wie der Bundesrat auf eine parlamentarische Anfrage antwortet. Vgl. Geschäftsdatenbank Curia Vista (2015). 15.4001 – Interpelation. US-Swiss Safe Harbor Framework. Die Personendaten wirklich schützen. URL http://www.parlament.ch/d/suche/seiten/geschaefte.aspx?gesch_id=20154001. Abruf: 20. November 2015.

5.2.1 Chilling Effects: Freiheit durch Anpassung?

Es existieren Studien [113], die vor und nach den Snowden-Enthüllunge nahelegen, dass sich Personen anders verhalten – mithin: selbst zensieren – wenn sie sich der ständigen Überwachung und der möglichen (sozialen) Sanktionsmöglichkeiten gewahr sind. Solche Effekte werden *Chilling Effects* genannt.

Chilling Effects der Massenüberwachung, die demnach als Folge der Snowden-Enthüllungen gemessen wurden, sind:

- Gemäss einer norwegischen Studie werden Websuchen in beschränkterem Umfang ausgeführt.
- In einer US-Studie wird nahegelegt, dass Personen aktiv versuchen, digitale Spuren zu verwischen oder Anonymisierungssoftware wie *Tor* zu verwenden.
- Hotline-Anrufe nehmen in der Post-Snowden-Zeit gemäss Erhebungen unterschiedlicher Verbände in den USA ab.
- US-Schriftsteller zensieren sich in bestimmten Themen selber: sowohl was ihre mündliche als auch schriftliche Äusserung betrifft.
- Eine US-Studie legt nahe, dass die Pressefreiheit abnehme, weil Informanten vorsichtiger seien.

Doch schon vor den Snowden-Enthüllungen legt dieselbe Quelle [113] Effekte offen, die sich bereits nach den 9/11-Anschlägen geäussert haben sollen; es handelt sich dabei um US-Studien:

- Minderheiten muslimischer Abstammung fühlen sich im Internet überwacht, insbesondere seit den 9/11-Anschlägen.
- Arbeitnehmende fühlen sich gestresst, wenn sie wissen, dass sie am Arbeitsplatz überwacht werden.
- Im sozialpsychologischen Experiment “Asch” wird darauf hingewiesen, dass unter Konformitätsbedingungen offensichtlich falsche Aussagen als richtig akzeptiert werden.

Soziolinguistisch betrachet stellen sich ebenfalls Fragen darüber, wie sich Menschen überhaupt sprachlich ausdrücken können und sollen, um der – unbekannten – Norm zu entsprechen, die verhindert, dass das “Raster” der Massenüberwachuchung greift. Anders als in Alltagssituationen, wo Regelverletzungen – entsprechend den geltenden sprachlichen Erwartungshaltungen – in Gesprächssituationen gegebenenfalls zu Gefühlen der Peinlichkeit führen können, kann eine Markeriung als “Terrorgefahr” im Rahmen einer Massenüberwachung ernsthafte Konsequenzen für den eigenen Lebensverlauf haben. [62, S.351ff.]

5.2.2 Anonymisierung und Obfuskation der Urheberschaft

Es existiert Software wie *Anonymouth*[134] [9, 106], die es erlaubt Texte zu anonymisieren oder doch zumindest wesentlich zu entstellen, indem stilistische Eigenschaften, andere und kürzere Wörter (mit geringeren globalen Auftretenshäufigkeiten) auf Basis textuellen Materials anderer Autoren übernommen werden. Dafür wird der eigene Text durch Trainingskorpora anderer Autoren obfuskiert, so dass Messmethoden forensischer Linguistik ungenauer werden.

Konkrete Praxis (autoren-)anonymisierender Kommunikation für den Voice-Bereich findet beim Hackerkollektiv *Anonymous* [54] statt: schon bei der ersten öffentlichen *Anonymous*-Aktion, der *Operation Chanology*, die sich gegen die *Scientology*-Sekte gerichtet hat, wurde zu Sprachtechnologie des Text-To-Speech gegriffen, um die Identifikation der Aktivisten anhand Verfahren der Stimmidentifikation zu erschweren.[135]

5.2.3 Anonymisierung und Verschlüsselung der Daten(-Wege)

Es sind der Massenüberwachung Grenzen setzbar: diese hängen mit dem Umstand zusammen, dass mit Einsatz bestimmter Soft- und Hardware, dem Überwachungskomplex (unter steigenden Kosten) erschwert wird, jeden Akteur beliebig zu überwachen. [12, 44, 129]

Andererseits kann der Einsatz von Verschlüsselungsverfahren für Geheimdienste Verdachtsmoment sein, im Sinne "konspirativen Verhaltens". Eine solche Argumentation bietet sich je mehr an, desto weniger Akteure verschlüsseln.

Gerade begünstigt durch die Snowden-Enthüllungen haben diverse Projekte die Initiative ergriffen, für mehr verschlüsselten Traffic zu sorgen. Ein spannendes Beispiel eines Projektes, das den Standard textbasierter Kommunikation auf anonymisierend und verschlüsselt abändern möchte, ist pretty Easy privacy oder p≡p.[136] Unter Einsatz einer Core-Engine, die mehrere anerkannte Verschlüsselungsverfahren bedient[137], indem anerkannte Crypto-Bibliotheken eingesetzt und Adaptoren für verschiedene Programmiersprachen zur Verfügung gestellt werden, wird eine plattformübergreifende Lösung geschaffen, die es ermöglicht auf diversen Kanälen und untereinander Nachrichten automatisch zu verschlüs-

[134] Vgl. Quellcode auf *github.com*: `https://github.com/psal/anonymouth`. Abruf: 12. November 2015.

[135] Noch immer aber könnte auf Grund der Textstruktur, welche dem TTS-System als Eingabe dient, möglicherweise auf die Autorenschaft geschlossen werden, falls diese nicht wie unter Kapitel 5.2.2 erläutert beispielsweise mit *Anonymouth* im Ursprung anonymisiert wird.

[136] Disclaimer: Ich arbeite selber in diesem Projekt verschiedentlich mit.

[137] Gemeint ist, dass Schritte, die üblicherweise End-User ausführen, um Schlüsselmaterial für eine vertrauliche Kommunikation zu erstellen (Keymanagement) und weitere Schritte, die ihn dazu befähigen, mit anderen End-Usern zuletzt verschlüsselt zu kommunizieren, in einer standardisierten Art und Weise automatisiert werden, dass diese folglich als Protokoll verschlüsselter Textkommunikation institutionalisiert werden können.

seln. Der Ansatz ist letztlich, die gesamte Textkommunikation zunehmend ins Darknet[138] zu verlagern, wo Akteure peer-to-peer miteinander kommunizieren, ohne zentrale dienstspezifische Vermittlungsstelle: durch die zusätzliche Verschlüsselung mit Verfahren wie OTR würden schliesslich bei den Geheimdiensten nicht nur punkto Inhaltsdaten, sondern auch Metadaten die Lichter ausgehen.

Zu beachten ist, dass selbst Verfahren der Ende-zu-Ende-Verschlüsselung nur in Fällen schützend wirken, wo der Datenzugriff zwischen den Akteuren erfolgt: kann gemäss den Ausführungen von 2.1.2 ein Endgerät kontrolliert und die Daten auf eine Überwachungsplattform ausgeleitet werden, ist zumindest eine Inhaltsüberwachung wieder sofort möglich.

Weiterhin bestehen für andere Formen der Kommunikation und Netznutzung Angriffsmethoden auf verschlüsselte Dienste wie *VPN*, *SSH* oder *Tor*, die im Allgemeinen als sicher gelten und doch unter gewissen (unwahrscheinlichen) Randbedingungen von der *NSA* und ihren Verbündeten umgangen werden können. [111]

5.3 Fazit

Die Ergebnisse zeigen auf, dass einige der Modelle besser als andere funktionieren, verdächtiges Material auf Basis einfacher Trainingskorpora in Textkollektionen zu finden: dies wenn (1) die Suchbegriffe spezifisch genug und (2) das Evaluationsmaterial in genügendem Umfang ausfällt, Treffer zu produzieren.

Sind die Suchbegriffe wenig spezifisch auf das Verdachtsmaterial ausgelegt, ist der Anteil von False-Positives naturgemäss höher, was folglich bedeutet, dass bei solchen Formen der Datenauswertung die Überwachung eine erhöhte Anzahl Menschen in den Verdacht rückt, einer bestimmten Gruppe anzugehören oder doch zumindest damit verbundene Aktivitäten beziehungsweise entsprechende Interessen zu hegen.

Gehen wir davon aus, dass heutige Geheimdienste Zugriff auf nahezu alle Formen unverschlüsselter Kommunikation haben und in vielen Fällen zusätzlich fähig sind, verschlüsselte Formen der Kommunikation zu umgehen, so kann von keinem Sparse-Data-Problem ausgegangen werden.

Es gelingt folglich mit sehr hoher Wahrscheinlichkeit Treffer zu produzieren, die ins "Raster" passen und somit Verdächtige zu generieren: sei es, weil solche bestimmte Webseiten ansurfen oder in ihrer Kommunikation die nötigen Begriffskombinationen führen. Genauso kann ins Raster fallen, wer "zufälligerweise" dieselben Suchbegriffe in Suchmaschinen eintippt,

[138] Angedacht ist GNUnet, ein offen spezifiziertes peer-to-peer-System, das anonymisierte und dezentralisierte Kommunikation zwischen diversen Knotenpunkten in einem Netzwerk erlaubt.
Vgl. Webseite GNUnet (2015). *GNU's Framework for Secure Peer-to-Peer Network*.
URL `https://gnunet.org/`. Abruf: 13. November 2015.

welche als Selektoren selber bestimmt sind, nach Sprache zu suchen, die von der Norm abweicht.

Wo ein erhebliches Willkürpotenzial besteht, ist in der Interpretation der Treffer gegeben, die anfallen. Fragen, die nicht ohne Weiteres von einem automatisch operierenden Überwachungssystem beantwortet werden können, sondern viel eher psychologische, soziologische, informationstechnische oder gar zufälligen Hintergrund haben können, sind beispielsweise – ohne abschliessend zu sein – Fragen der folgenden Art:

- Wieso hat jemand eine bestimmte Webseite aufgerufen?
- Ist die Person, von deren Anschluss oder Webbrowser-Kennung aus die Webseite aufgerufen wurde, auch dieselbe, die die Webseite aufgerufen hat?[139]
- Ist die Person wirklich eine verdächtige Person oder recherchiert sie bloss für einen Artikel oder eine Forschungsarbeit?

Diese Fragen auch nur ansatzweise zu beantworten, erfordert einerseits eine umfassende Form der Überwachung jedweder Aktivitäten einer konkreten Person und hat andererseits das Potenzial, es erforderlich zu machen, auch das gesamte Umfeld einer Person mit zu überwachen, um beispielsweise harmlose Forschungs- oder Studierendengruppen (etwa an Universitäten), Journalisten oder etwa (strafrechtliche) Ermittler als Verdächtige auszuschliessen, die sich beispielsweise mit "politischen Randgruppen" beschäftigen. Diese Praxis birgt die Gefahr, ausufernden Charakter anzunehmen, in einer Form, dass die Überwachung letztlich nicht nur oberflächlich, sondern auch invasiv auf diverse Bevölkerungskreise ausgedehnt wird. Exemplarisch hierfür kann der Fall von Andrej Holm gemäss Abschnitt 2.1.3.2 stehen, der mithin im benachbarten Deutschland in Isolationshaft kam: eines der wichtigen Anfangsverdachtsmomente stellte eine behördliche *Google*-Suche dar, die Treffer mit seinen Texten zu Tage förderte.

Zusätzlich fällt rasch auf, dass eine zweite Dimension der Willkür dem System der Generierung entsprechender Selektoren direkt inhärent ist. Je nach eingesetztem Trainingsmaterial resultierten (leicht) andere Selektoren. Das heisst gleichzeitig, dass je nach Umfang und Vorverarbeitung des Trainingsmaterials andere Selektoren erzeugt werden, die zu anderen Treffern und entsprechend anderen Verdächtigen führen können.

Zudem liegt die Missbrauchsgefahr solcher Systeme auf der Hand: ein System, das mit Selektoren für "politischen Extremismus" trainiert wurde, kann durch Einsatz anderen Textmaterials und gegebenfalls leicht differierender Logik zur Selektorgenerierung für diplomatische Ausspähung oder Wirtschaftsspionage eingesetzt werden, wie dies ganz offensichtlich gemäss Snowden-Enthüllungen von Seiten der USA oder – gemäss laufender parlamentarischer Untersuchungen – seitens Deutschland en vogue ist.

[139]Schadsoftware oder (unwissentlich) laufende Crawler-Software, wie solche die bei *YaCy* im Einsatz ist, können für Webseitenabrufe sorgen.

Das Grundproblem zuletzt bleibt, dass wer mit Suchbegriffen nach der Nadel im Heuhaufen sucht, immer False-Positives fördert. Auf der anderen Seite sind auch False-Negatives vorhanden, wenn sich Akteure verschlüsselnder oder anonymisierender Technik bedienen, oder die eingesetzen Selektoren auf (grund-)falschen Annahmen beruhen: sie also irrelevante Treffer erzeugen.

Allen Personen mit technischem Verständnis wird bewusst sein, dass ohne besondere Vorkehrungen zu treffen, die elektronische Kommunikation nachweislich der vollständigen Überwachung untersteht. Zu keinem Zeitpunkt kann allgemein den Menschen hingegen bewusst sein, dass konkret ihre Daten für eine spätere Verarbeitung oder genauere Analyse aussortiert werden. Das mangelnde Wissen um die Funktionsweise dieser Systeme, die im Einsatz befindlichen konkreten Selektoren[140] und daraus resultierende Treffer werfen Fragen auf, ob solche Formen der Überwachung in Rechtsstaaten mit demokratischer Verfassung und Bekenntnis zu Menschenrechten tragbar sind: dies umso mehr, wird zur Kenntnis genommen, dass die meisten Staatsvölker oder gar "die" globale Gesellschaft nie über ein solches Ausmass der Überwachung – informiert – abgestimmt und solcher Praxis zumindest den Hauch einer Legitimität zugesprochen haben.

Doch statt die nötig dringende Diskussion über die laufende Praxis zu führen, werden Überwachungsgesetze sowohl hierzulande als auch im benachbarten Ausland (etwa in Deutschland oder Frankreich) ausgebaut, offenbar im Aufrüstungseifer den "Vorbildern" des Überwachungskomplexes der *FVEY* nahe zu kommen und genauso erhebliche Kapazitäten zur Überwachung lokalen wie globalen Datenverkehrs aufzubauen.

Am Ende angekommen, bleibt die Hoffnung, dass mit dieser Arbeit ein wertvoller Anstoss zur Aufklärung und Diskussion der gegenwärtig und zukünftig (weiter) möglichen lokalen wie globalen Massenüberwachung gegeben ist.

[140] Oder nur schon ihre (Entstehungs-)Natur oder die Gesamtheit ihrer Kategorien.

Glossar

An dieser Stelle sind einige wichtige Begriffe aufgeführt, die in der Arbeit häufig vorkommen. Ein umfassenderes Glossar im Zusammenhang mit den Snowden-Enthüllungen findet sich bei *BBC*. [45]

Inhaltsdaten Inhaltsdaten der Kommunikation sind grundsätzlich der "Payload" oder die Kernbotschaft, die bei Kommunikationsdaten bestehen: das können statt nur Webseiten-URLs oder Zeiten des Abrufs, der konkrete Inhalt der Seite sein, statt nur der Dateiname einer Datei. Bei der Massenüberwachung nach Inhaltsdaten müssen grundsätzlich alle Daten durchsucht werden: auch diejenigen, die dem Metadaten-Bereich zufallen.

Massenüberwachung Als Überwachungsparadigma zeichnet sich Massenüberwachung dadurch aus, dass anstatt zielgerichtet konkrete Verdächtige zu überwachen, alle dem Generalverdacht unterstellt werden, verdächtigs Verhalten aufzuweisen. Merkmale von Massenüberwachung sind, dass sie anlasslos (ständig) erfolgt und verdachtsunabhängig alle betrifft.

Metadaten Metadaten sind technische Kommunikationsmerkmale der Kommunikation wie E-Mail-, IP-Adressen, aber auch Chat-Nicknames und in E-Mails Betreffszeilen. Die Auswertung von Metadaten alleine können bereits ein umfassendes Bild der Kontaktnetzwerke und Interessen (beispielsweise beim Websurfen) einer Person liefern.

Onyx Ein Schweizer System zur Massenüberwachung von Funkverbindungen, das vom Zentrum für Elektronische Operationen ZEO betrieben wird. Auftraggeber sind Militär und der Schweizer Geheimdienst NDB.

PRISM Ein Programm von FBI und NSA, das dazu genutzt wird, Daten von Online-Providern wie unter anderem Apple, Google oder Facebook auszuleiten und beispielsweise von XKeyscore durchsuchbar zu machen.

Selektor Ein Suchbegriff, der sowohl formaler oder technischer, als auch inhaltlicher Art sein kann. Erstere sind "Strong-", zweitere "Soft"-Selektoren. Formale Selektoren sind Selektoren, welche technische Kommunikationsmerkmale wie E-Mail-, IP-Adressen oder Telefon-Nummern betreffen. Grenzfälle sind gegeben, wenn Named Entities wie Personnamen oder Organizationsbezeichnungen als Seleltkoren im Einsatz sind. Diese können sowohl darauf aus sein, Metadaten in Datenströmen als auch eigentliche Inhaltsdaten zu finden.

Tempora Ein Programm vom britischen Geheimdienst GCHQ, das Daten aus Unterseekabeln ausleitet und sie zum Beispiel mittels der NSA-Suchmaschine XKeyscore durchsuchbar macht.

XKeyscore Eine NSA-Suchmaschine, die es erlaubt, in privaten Datenströmen, mit beliebigen Selektoren oder auch komplexeren (verzweigten) "Ausdrücken" zu suchen. Es können beispielsweise fingerprints in C++ geschrieben werden, die es ermöglichen, die Datenströme nur von gewissen Geräten – beispielsweise Apple iPhones – zu durchsuchen oder Treffer von Systemen weltweit anzuzeigen, die bestimmte Sicherheitslücken haben.

Quellenverzeichnis

[1] Peggy Becker und Dick Holdsworth. Scientific and Technological Options Assessment STOA. Development of Surveillance Technologies and Risk of Abuse of Economic Information. *Europäisches Parlament. PE 168.184/Vol 1/5/EN*, Dezember 1999. URL http://www.europarl.europa.eu/RegData/etudes/etudes/join/1999/168184/DG-4-JOIN_ET(1999)168184_EN.pdf.

[2] Kai Biermann. BND-Spionage ... oder Merkel hat gelogen. *Zeit Online*, 25. November 2015. URL http://www.zeit.de/politik/deutschland/2015-11/bnd-selektoren-nsaua-merkel. Abruf: 26. November 2015.

[3] Kai Biermann. NSA-Programm XKeyscore. Diese Spähsoftware findet jedes Passwort. *Zeit Online*, 27. August 2015. URL http://www.zeit.de/digital/datenschutz/2015-08/bfv-verfassungsschutz-was-kann-xkeyscore/komplettansicht. Abruf: 23. November 2015.

[4] Kai Biermann und Patrick Beuth. Bundesnachrichtendienst: Was sind eigentlich Selektoren? *Zeit Online*, 24. April 2015. URL http://www.zeit.de/digital/datenschutz/2015-04/bundesnachrichtendienst-bnd-nsa-selektoren-eikonal. Abruf: 30. November 2015.

[5] Volker Birk. Was die Tagesschau in den Nutzerkommentaren herausfiltert. *>b's weblog*, 27. November 2015. URL http://blog.fdik.org/2015-11/s1448619468. Abruf: 28. November 2015.

[6] Anna Biselli. Textanalyse. Unter Generalverdacht durch Algorithmen. *golem.de (Online)*, 19. Februar 2014. URL http://www.golem.de/news/textanalyse-unter-generalverdacht-durch-algorithmen-1402-104637.html. Abruf: 22. November 2015.

[7] Anna Biselli. IMSI, IMEI, SIP: Selektoren-Gutachter Graulich verheddert sich im Technik-Dschungel. *Netzpolitik.org*, 4. November 2015. URL https://netzpolitik.org/2015/imsi-imei-sip-selektoren-gutachter-graulich-verheddert-sich-im-technik-dschungel/. Abruf: 30. November 2015.

[8] David M. Blei, Andrew Y. Ng, und Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3: 993–1022, 1. März 2003. ISSN 1532-4435. URL http://machinelearning.wustl.edu/mlpapers/paper_files/BleiNJ03.pdf. Abruf: 19. November 2015.

[9] Michael Brennan, Sadia Afroz, und Rachel Greenstadt. Deceiving Authorship Detection. Tools to Write Anonymously & Current Trends in Adversarial Stylometry. Folien, 28. Chaos Communication Congress 28C3, 29. Dezember 2011. URL https://events.ccc.de/congress/2011/Fahrplan/attachments/2019_28C3-authorship.pdf. Abruf: 22. November 2015.

[10] Nina Brink und Constanze Kurz. Grundrechte-Report 2015: Massenüberwachung ist nicht abstrakt. Constanze Kurz im Gespräch mit Nana Brink. *Deutschlandradio Kultur (Online)*, 22. Mai 2015. URL http://www.deutschlandradiokultur.de/grundrechte-report-2015-massenueberwachung-ist-nicht.1008.de.html?dram:article_id=320569. Abruf: 19. November 2015.

[11] Noah Bubenhofer und Joachim Scharloth. Maschinelle Textanalyse im Zeichen von Big Data und Data-driven Turn – Überblick und Desiderate. *Zeitschrift für germanistische Linguistik*, 43(1):1–26, März 2015. ISSN 0301-3294.

[12] Philip Bump. So, You Want to Hide from the NSA? Your Guide to the Nearly Impossible. *The Wire (Online)*, 9. Juli 2013. URL http://www.thewire.com/technology/2013/07/so-you-want-hide-nsa-your-guide-nearly-impossible/66942/. Abruf: 26. November 2015.

[13] Deutscher Bundestag. 1. Untersuchungsausschuss ("NSA"). URL http://www.bundestag.de/bundestag/ausschuesse18/ua/1untersuchungsausschuss. Abruf: 27. November 2015.

[14] Duncan Campbell. Somebody's listening. *New Statesman*, pages 10–12, 12. August 1988. URL http://web.archive.org/web/20130420093650/http://duncan.gn.apc.org/echelon-dc.htm. Abruf: 23. November 2015.

[15] Cryptome.org. National Intelligence Priorities Framework NIPF. Chart. Composited from different timeframes. April 2014. URL https://cryptome.org/2014/04/nipf-v4.pdf. Abruf: 27. November 2015.

[16] Krzysztof Dorosz. D9.6. System for enhanced search: tool for pattern based information retrieval. *Paper im Rahmen des 7. Europäischen Forschungrahmenprogramms FP7 (Projekt INDECT 218086)*, 11. Januar 2012. URL http://www.indect-project.eu/files/deliverables/public/deliverable-9.6/at_download/file. Abruf: 22. November 2015.

[17] Krzysztof Dorosz, Michał Korzycki, und Wieslaw Lubaszewski. D4.4. System for Enhanced Search: A Tool for Pattern Based Information Retrieval. *Paper im Rahmen des 7. Europäischen Forschungrahmenprogramms FP7 (Projekt INDECT 218086)*, 30. Dezember 2009. URL http://www.indect-project.eu/files/deliverables/public/deliverable-4.4/at_download/file. Abruf: 22. November 2015.

[18] Krzysztof Dorosz, Michał Korzycki, und Wiesław Lubaszewski. D9.16 Combined Tool for Enhanced Search. *Paper im Rahmen des 7. Europäischen Forschungrahmenprogramms FP7 (Projekt INDECT 218086)*, 17. August 2011. URL http://www.indect-project.eu/files/deliverables/public/deliverable-9.16/at_download/file. Abruf: 27. November 2015.

[19] DuckDuckGo. Fact sheet. *DuckDuckGo. Community Platform*, . URL https://duck.co/help/company/fact-sheet. Abruf: 28. November 2015.

[20] DuckDuckGo. Sources. *DuckDuckGo. Community Platform*, . URL https://duck.co/help/results/sources. Abruf: 28. November 2015.

[21] Sarah Ebling, Joachim Scharloth, Tobias Dussa, und Noah Bubenhofer. Gibt es eine Sprache des politischen Extremismus? [Preprint]. In Frank Liedtke, editor, *Sprache, Politik, Partizipation*. Hempen, Bremen, 2012. URL http://www.scharloth.com/publikationen/scharloth_extremismus_preprint.pdf. Abruf: 30. November 2015.

[22] Urs Paul Engeler. Abhörsystem: Was sagen Sie jetzt? 18. März 2005. URL http://www.weltwoche.ch/ausgaben/2005-10/artikel-2005-10-was-sagen-sie-je.html. Abruf: 30. November 2015.

[23] Freedom for Georges Ibrahim Abdallah. Who Is Georges Abdallah: The longest-held political prisoner of the European continent. URL http://www.freegeorges.org/who-is-georges-abdallah/. Abruf: 19. November 2015.

[24] Dan Froomkin. The Computers Are Listening. How the NSA Converts Spoken Words Into Searchable Text. *The Intercept*, 5. Mai 2015. URL https://theintercept.com/2015/05/05/nsa-speech-recognition-snowden-searchable-text/. Abruf: 22. November 2015.

[25] Bündnis für die Einstellung der §129(a) Verfahren. Offener Brief an die Generalbundesanwaltschaft. Offener Brief an die Generalbundesanwaltschaft gegen die Kriminalisierung von kritischer Wissenschaft und politischem Engagement. 15. August 2007. URL http://einstellung.so36.net/de/offenerbrief. Abruf: 25. November 2015.

[26] Bündnis für die Einstellung der §129(a) Verfahren. *Das zarte Pflänzchen der Solidarität gegossen*. edition assemblage, Münster, 2011. ISBN 978-3-942885-00-3. URL https://einstellung.so36.net/files/buch_einstellung_129_mg_verfahren.pdf. Abruf: 19. November 2015.

[27] Carlos Gacimartín, Alberto José Hernández, Manuel Urueña, und David Larrabeiti. On detecting Internet-based criminal threats with XplicoAlerts: Current design and next steps. *Paper im Rahmen des 7. Europäischen Forschungrahmenprogramms FP7 (Projekt INDECT 218086)*, 2010. URL http://www.it.uc3m.es/muruenya/papers/MCSS10XplicoAlerts.pdf. Abruf: 22. November 2015.

[28] Ryan Gallagher und Glenn Greenwald. How the NSA Plans to Infect "Millions" of Computers with Malware. *The Intercept*, 12. März 2014. URL https://theintercept.com/2014/03/12/nsa-plans-infect-millions-computers-malware/. Abruf: 22. November 2015.

[29] Bogdan Gliwa, Anna Zygmunt, und Jarosław Koźlak. Analysis of roles and groups in blogosphere. *Paper im Rahmen des 7. Europäischen Forschungrahmenprogramms FP7 (Projekt INDECT 218086)*, 19. Juni 2013. URL http://arxiv.org/pdf/1306.4598v1.pdf. Abruf: 22. November 2015.

[30] Geschäftsprüfungsdelegation der Eidgenössischen Räte GPDel. Satellitenaufklärungssystem des Eidgenössischen Departements für Verteidigung, Bevölkerungsschutz und Sport (Projekt «Onyx»). Bericht der Geschäftsprüfungsdelegation der Eidgenössischen Räte. 10. November 2003. URL http://www.parlament.ch/d/dokumentation/berichte/berichte-delegationen/berichte-der-geschaeftspruefungsdelegation/Documents/ed-pa-gpd-onyx-d.pdf. Abruf: 28. November 2015.

[31] Geschäftsprüfungsdelegation der Eidgenössischen Räte GPDel. Mitbericht der Geschäftsprüfungsdelegation zum NDG (14.022). 22. April 2014. URL http://www.parlament.ch/d/organe-mitglieder/delegationen/geschaeftspruefungsdelegation/nachrichtendienstgesetz/Documents/gpdel-mitbericht-2014-04-22-d.PDF. Abruf: 30. November 2015.

[32] Geschäftsprüfungsdelegation der Eidgenössischen Räte GPDel und Geschäftsprüfungskommissionen der Eidgenössischen Räte GPK. Jahresbericht 2007 der Geschäftsprüfungskommissionen und der Geschäftsprüfungsdelegation der Eidgenössischen Räte. 08.004. 25. Januar 2008. URL https://www.admin.ch/opc/de/federal-gazette/2008/5061.pdf. Abruf: 30. November 2015.

[33] Geschäftsprüfungsdelegation der Eidgenössischen Räte GPDel und Geschäftsprüfungskommissionen der Eidgenössischen Räte GPK. Jahresbericht 2008 der Geschäftsprüfungskommissionen und der Geschäftsprüfungsdelegation der Eidgenössischen Räte. 08.004. 23. Januar 2009. URL https://www.admin.ch/opc/de/federal-gazette/2009/2575.pdf. Abruf: 30. November 2015.

[34] Geschäftsprüfungsdelegation der Eidgenössischen Räte GPDel und Geschäftsprüfungskommissionen der Eidgenössischen Räte GPK. Jahresbericht 2011 der Geschäftsprüfungskommissionen und der Geschäftsprüfungsdelegation der Eidgenössischen Räte. 12.004. 27. Januar 2012. URL https://www.admin.ch/opc/de/federal-gazette/2012/6783.pdf. Abruf: 30. November 2015.

[35] Geschäftsprüfungsdelegation der Eidgenössischen Räte GPDel und Geschäftsprüfungskommissionen der Eidgenössischen Räte GPK. Jahresbericht 2012 der Geschäftsprüfungskommissionen und der Geschäftsprüfungsdelegation der Eidgenössischen Räte. 13.004. 24. Januar 2013. URL http://www.parlament.ch/d/dokumentation/berichte/berichte-aufsichtskommissionen/geschaeftspruefungskommission-GPK/berichte-2013/Documents/jahresbericht-gpk-ns-2012-d.pdf. Abruf: 30. November 2015.

[36] Kurt Graulich. Nachrichtendienstliche Fernmeldeaufklärung mit Selektoren in einer transnationalen Kooperation. Prüfung und Bewertung von NSA-Selektoren nach Massgabe des Beweisbeschlusses BND-26. *Bericht im Rahmen des 1. Untersuchungsausschusses der 18. Wahlperiode des Deutschen Bundestages*, 23. Oktober 2015. URL https://www.bundestag.de/blob/393598/b5d50731152a09ae36b42be50f283898/mat_a_sv-11-2-data.pdf. Abruf: 30. November 2015.

[37] Michael Götschenberg. BND hörte deutschen Diplomaten ab. *Rundfunk Berlin-Brandenburg rbb (Online)*, 11. November 2015. URL http://www.rbb-online.de/politik/beitrag/2015/11/bnd-selektoren-diplomat-abgehoert-aussenminister-frankreich.html. Abruf: 30. November 2015.

[38] Carlos Hanimann. «Operation Transit»: Spionagebefehl aus dem Kanzleramt. *Wochenzeitung WOZ (Online)*, 22. Oktober 2015. URL https://www.woz.ch/1543/operation-transit/spionagebefehl-aus-dem-kanzleramt. Abruf: 13. November 2015.

[39] Dirk Helbling, Bruno S. Frey, Gerd Gigerenzer, Ernst Hafen, Michael Hagner, Yvonne Hofstetter, Jeroen van den Hoven, Roberto V. Zicari, und Andrej Zwitter. IT-Revolution: Digitale Demokratie statt Datendiktatur ("Das Digital Manifest"). *Spektrum der Wissenschaft (Online)*, 12. November 2015. URL http://www.spektrum.de/news/wie-algorithmen-und-big-data-unsere-zukunft-bestimmen/1375933. Abruf: 18. November 2015.

[40] Martin Holland. 30C3: Überwachungsalgorithmen und die "Radikalität" von Fefes Blog. *heise online*, 28. Dezember 2013. URL http://www.heise.de/newsticker/meldung/30C3-Ueberwachungsalgorithmen-und-die-Radikalitaet-von-Fefes-Blog-2072600.html. Abruf: 27. November 2015.

[41] Michael Isikoff. NSA program stopped no terror attacks, says White House panel member. *NBC News (Online)*, 20. Dezember 2013. URL http://www.nbcnews.com/news/other/nsa-program-stopped-no-terror-attacks-says-white-house-panel-f2D11783588. Abruf: 18. November 2015.

[42] Andrea Jonjic. USA: Massenüberwachung konnte auch vor Snowden-Leaks keine Terrorangriffe vereiteln. *Netzpolitik.org*, 18. November 2015. URL https://netzpolitik.org/2015/usa-massenueberwachung-konnte-auch-vor-snowden-leaks-keine-terrorangriffe-vereiteln/. Abruf: 18. November 2015.

[43] Suraj Jung Pandey und Krzysztof Dorosz. D4.11 Specification of methods for mining and detecting suspicious websites. *Paper im Rahmen des 7. Europäischen Forschungrahmenprogramms FP7 (Projekt INDECT 218086)*, Juli 2012. URL http://www.indect-project.eu/files/deliverables/public/deliverable-4.11/at_download/file. Abruf: 22. November 2015.

[44] V. Kashyap und P. Boominathan. Privacy Shielding against Mass Surveillance. *International Journal of Engineering Trends and Technology IJETT*, 8(2), Februar .

[45] Leo Kelion. NSA-GCHQ Snowden leaks: A glossary of the key terms. *Britisch Broadcasting Corporation BBC*, 28. Januar 2014. URL http://www.bbc.com/news/technology-25085592. Abruf: 30. November 2015.

[46] McCarthy Kieren. What are those words that trigger Echelon? We'll tell you. *The Register*, 31. Mai 2001. URL http://www.theregister.co.uk/2001/05/31/what_are_those_words/. Abruf: 28. November 2015.

[47] Ioannis P. Klapaftis. Report on current state-of-the-art methods for relationship mining. *Paper im Rahmen des 7. Europäischen Forschungrahmenprogramms FP7 (Projekt INDECT 218086)*, 30. Oktober 2009. URL http://www.indect-project.eu/files/deliverables/public/INDECT_Deliverable_D4.2_v20091030.pdf/at_download/file. Abruf: 22. November 2015.

[48] Ioannis P. Klapaftis. Report on methodology for applying existing machine learning methods for behavioural profiling. *Paper im Rahmen des 7. Europäischen Forschungrahmenprogramms FP7 (Projekt INDECT 218086)*, Juli 2011. URL http://www.indect-project.eu/files/deliverables/public/deliverable-4.7/at_download/file. Abruf: 22. November 2015.

[49] Ioannis P. Klapaftis. Novel algorithms for relationship mining including comparison with existing methods (Novel algorithms for relationship mining including comparison with existing methods indicated in D4.2). *Paper im Rahmen des 7. Europäischen Forschungrahmenprogramms FP7 (Projekt INDECT 218086)*, Mai 2011. URL http://www.indect-project.eu/files/deliverables/public/deliverable-9.14/at_download/file. Abruf: 22. November 2015.

[50] Ioannis P. Klapaftis. D4.5 Novel algorithms for relationship mining including comparison with existing methods indicated in D4.2. *Paper im Rahmen des 7. Europäischen Forschungrahmenprogramms FP7 (Projekt INDECT 218086)*, 21. April 2011. URL http://www.indect-project.eu/files/deliverables/public/deliverable-4.5/at_download/file. Abruf: 22. November 2015.

[51] Ioannis P. Klapaftis und Suraj Jung Pandey. D4.9 Novel algorithms for behavioural profiling and comparison with baseline systems developed in 4.3. *Paper im Rahmen des 7. Europäischen Forschungrahmenprogramms FP7 (Projekt INDECT 218086)*, Juni 2012. URL http://www.indect-project.eu/files/deliverables/public/deliverable-4.9/at_download/file. Abruf: 22. November 2015.

[52] Klapaftis, Ioannis P. and Nagy, Zoltán and Johanning, Nils. Report on current state-of-the-art methods for relationship mining. WP9. D.9.5. *Paper im Rahmen des 7. Europäischen Forschungrahmenprogramms FP7 (Projekt INDECT 218086)*, 31. Oktober 2009. URL http://www.indect-project.eu/files/deliverables/public/INDECT_Deliverable_D9.5_v20091031.pdf/at_download/file. Abruf: 22. November 2015.

[53] Thomas Knellwolf. Hintergrund: Britische Spione lesen Schweizer E-Mails. *Tages-Anzeiger TA (Online)*, 26. Juni 2013. URL http://www.tagesanzeiger.ch/ausland/Der-Freund-hoert-mit/story/31982299. Abruf: 28. November 2015.

[54] Markus Kompa. Anonymous: Meinungsfreiheit hat (k)einen Namen. *Telepolis*, 28. August 2014. URL http://www.heise.de/tp/artikel/42/42596/1.html. Abruf: 27. November 2015.

[55] Michał Korzycki und Wiesław Lubaszewski. D9.15 System for Enhanced Search: A Tool for Associative Information Retrieval. *Paper im Rahmen des 7. Europäischen Forschungrahmenprogramms FP7 (Projekt INDECT 218086)*, Mai 2011. URL http://www.indect-

project.eu/files/deliverables/public/deliverable-9.15/at_download/file. Abruf: 27. November 2015.

[56] Jarosław Koźlak und Anna Zygmunt. Agent-based modelling of social organisations. *Paper im Rahmen des 7. Europäischen Forschungrahmenprogramms FP7 (Projekt INDECT 218086)*, 25. März 2013. URL http://arxiv.org/pdf/1303.6091v1.pdf. Abruf: 22. November 2015.

[57] Constanze Kurz. NSA-Abhörskandal: So wichtig bin ich doch nicht. *Frankfurter Allgemeine Zeitung FAZ (Online)*, 13. Juli 2015. URL http://www.faz.net/aktuell/feuilleton/aus-dem-maschinenraum/akzeptieren-politiker-die-bespitzelung-als-norm-13699092.html?printPagedArticle=true. Abruf: 15. November 2015.

[58] Constanze Kurz. Geheimdienstkooperation: Fünf Zimmer, Küche, Selektor. *Frankfurter Allgemeine Zeitung FAZ (Online)*, 2. November 2015. URL http://www.faz.net/aktuell/feuilleton/aus-dem-maschinenraum/kurt-graulichs-vorwuerfe-gegen-die-nsa-13888040.html. Abruf: 30. November 2015.

[59] Constanze Kurz, Frank Rieger, Harald Staun, und Sascha Lobo. Handreichung zum Thema Hacken: Kein System ist sicher. *Frankfurter Allgemeine Zeitung FAZ (Online)*, 12. Juli 2015. URL http://www.faz.net/aktuell/feuilleton/debatten/die-digital-debatte/eine-handreichung-zum-thema-hacken-13658599.html?printPagedArticle=true. Abruf: 14. November 2015.

[60] J. Richard Landis und Garry G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, März 1977. URL http://www.dentalage.co.uk/wp-content/uploads/2014/09/landis_jr__koch_gg_1977_kappa_and_observer_agreement.pdf. Abruf: 28. November 2015.

[61] Antoni T. Ligęza, Weronika Adrian, Kaczor Krzysztof, Grzegorz J. Nalepa, Przemysław Ciężkowski, , Maciej Żywioł, und Paweł Grzesiak. D9.30 Web System for Citizen Provided Information, Automatic Knowledge Extraction, Knowledge Management and GIS Integration. *Paper im Rahmen des 7. Europäischen Forschungrahmenprogramms FP7 (Projekt INDECT 218086)*, 30. September 2012. URL http://www.indect-project.eu/files/deliverables/public/deliverable-9.30/at_download/file. Abruf: 27. November 2015.

[62] Angelika Linke, Markus Nussbaumer, und Paul R. Portmann. *Studienbuch Linguistik. Ergänzt um ein Kapitel "Phonetik / Phonologie" von Urs Willi. 5., erweiterte Auflage*. Max Niemeyer Verlag, Tübingen, 2004. ISBN 3-484-31125-5. Mit Ergänzungen von Simone Berchtold, Martin Businger, Jürg Fleischer, Franziska Gugger, Stefan Hauser, Jacqueline Holzer, Martin Luginbühl, Daniela Macher, Anna-Katharina Pantli, Joachim Scharloth, Jürgen Spitzmüller, Christa Stocker, Rebekka Studier.

[63] Schascha Lobo. Ausweitung der Überwachung: Geheimdienste lesen nicht mal Zeitung. *Der Spiegel (Online)*, 25. November 2015. URL http://www.spiegel.de/netzwelt/web/sascha-lobo-ueber-die-irrationale-ausweitung-der-ueberwachung-a-1064508.html. Abruf: 25. November 2015.

[64] Ewen MacAskill, Julian Borger, Nick Hopkins, und James Ball. Mastering the internet: how GCHQ set out to spy on the world wide web. Project Tempora – the evolution of a secret programme to capture vast amounts of web and phone data. *The Guardian (Online)*, 21. Juni 2013. URL http://www.theguardian.com/uk/2013/jun/21/gchq-mastering-the-internet. Abruf: 18. November 2015.

[65] Hernani Marques. INDECT-Forschungsergebnisse der Europäischen Union. Chancen und Gefahren der Möglichkeiten zur zentralisierten und transnationalen Überwachung. Seminararbeit, Soziologisches Institut, Universität Zürich, 15. Mai 2014. URL

https://www.ccczh.ch/images/a/a4/SOZ_140515--marques.hernani-indect4web.pdf. Abruf: 22. November 2015.

[66] Morgan Marquis-Boire, Glen Greenwald, und Micah Lee. XKEYSCORE: NSA's Google for the World's Private Communications. *The Intercept*, 1. Juli 2015. URL https://theintercept.com/2015/07/01/nsas-google-worlds-private-communications/. Abruf: 30. November 2015.

[67] Simon Marti. Ösis behaupten: NSA-Attacke auf die Swisscom! *Blick (Online)*, 21. Mai 2015. URL http://www.blick.ch/news/politik/oesis-behaupten-nsa-attacke-auf-die-swisscom-id3782784.html. Abruf: 13. November 2015.

[68] Georg Mascolo und John Goetz. BND: Die Überwachungsfabrik. *Süddeutsche Zeitung SZ (Online)*, 1. Mai 2015. URL http://www.sueddeutsche.de/politik/bnd-die-ueberwachungsfabrik-1.2460526. Abruf: 30. November 2015.

[69] Andre Meister. Live-Blog aus dem Geheimdienst-Untersuchungsausschuss: "Bis 2013 hätte der BND EU-Kommissare wie Oettinger überwacht". *Netzpolitik.org*, 10. September 2015. URL https://netzpolitik.org/2015/live-blog-aus-dem-geheimdienst-untersuchungsausschuss-bnd-mitarbeiter-zu-selektoren-verizon-zu-operation-glotaic/#zeuge1. Abruf: 30. November 2015.

[70] Korzyck Michał und Wiesław Lubaszewski. D4.8. Combined tool for enhanced search. *Paper im Rahmen des 7. Europäischen Forschungrahmenprogramms FP7 (Projekt INDECT 218086)*, 17. August 2011. URL http://www.indect-project.eu/files/deliverables/public/deliverable-4.8/at_download/file. Abruf: 22. November 2015.

[71] Markus H. F. Mohler. Nachrichtendienstgesetz: Ein echtes Staatsschutzgesetz sieht anders aus. *Oltner Tagblatt OT (Online)*, 18. Mai 2015. URL http://www.oltnertagblatt.ch/kommentare/ein-echtes-staatsschutzgesetz-sieht-anders-aus-129148449. Abruf: 28. November 2015.

[72] Daniel Mossbrucker. Digitale Informantenschutzrechte: Was offenbaren hinterlassene digitale Daten über die Umstände einer journalistischen Recherche und wie sind Journalisten vor einem Zugriff auf diese Daten durch Ermittlungsbehörden rechtlich geschützt? Bachelorarbeit, Technische Universität Dortmund, 15. August 2015. Mit Sperrvermerk versehene unpublizierte Arbeit: vgl. für Auszüge Anhang.

[73] Ellen Nakashima und Barton Gellman. National Security. Court gave NSA broad leeway in surveillance, documents show. *The Washington Post (Online)*, 30. Juni 2014. URL https://www.washingtonpost.com/world/national-security/court-gave-nsa-broad-leeway-in-surveillance-documents-show/2014/06/30/32b872ec-fae4-11e3-8176-f2c941cf35f1_story.html. Abruf: 22. November 2015.

[74] Nachrichtendienst des Bundes NDB. Sicherheit Schweiz: Jahresbericht 2010 des Nachrichtendienstes des Bundes. *Eidgenössisches Departement für Verteidigung, Bevölkerungsschutz und Sport*, 2010. URL http://www.vbs.admin.ch/internet/vbs/de/home/documentation/publication/snd_publ.parsys.5549.downloadList.35210.DownloadFile.tmp/ndbjahresbericht2010d.pdf. Abruf: 15. November 2015.

[75] Nachrichtendienst des Bundes NDB. Sicherheit Schweiz: Jahresbericht 2015 des Nachrichtendienstes des Bundes. *Eidgenössisches Departement für Verteidigung, Bevölkerungsschutz und Sport*, 2015. URL http://www.vbs.admin.ch/internet/vbs/de/home/documentation/publication/snd_publ.parsys.41649.downloadList.22907.DownloadFile.tmp/lageberichtndbd.pdf. Abruf: 15. November 2015.

[76] Nikolaj Nielsen. UK confirms it bulk-collected nationals' data. *EUobserver (Online)*, 5. November 2015. URL https://euobserver.com/justice/130973. Abruf: 22. November 2015.

[77] Center for Content Extraction) NSA, National Security Agency (Human Language Technology. Content Extraction Analytics SIGDEV End-to-End Demo. *American Civil Liberties Union ACLU*, 21. Mai 2009. URL https://www.aclu.org/sites/default/files/field_document/Content%20Extraction%20Analytics.pdf. Abruf: 28. November 2015.

[78] National Security Agency NSA. National Intelligence Program budget for fiscal year 2013. Research & Technology. Human Language Technology Research. *Eyeing the Five*, 5. Mai 2015 [1. Februar 2012] . URL https://fveydocs.org/document/hlt-research-bb-p360-364/. Abruf: 28. November 2015.

[79] National Security Agency NSA. How Is Human Language Technology (HLT) Progressing? *Eyeing the Five*, 5. Mai 2015 [6. September 2011] . URL https://fveydocs.org/document/hlt-progressing/. Abruf: 28. November 2015.

[80] National Security Agency NSA. United States SIGINT System January 2007 Strategic Mission List. *cryptome.org (Online)*, April 2014 [Januar 2007] . URL https://cryptome.org/2014/09/nsa-strategic-mission-list.pdf. Abruf: 22. November 2015.

[81] National Security Agency NSA. "writing xks fingerprints". 1. Juli 2015 [1. November 2010] . URL https://fveydocs.org/document/writing-xks-fingerprints/. Abruf: 30. November 2015.

[82] National Security Agency NSA. The Unofficial XKEYSCORE User Guide. *The Intercept*, 1. Juli 2015 [8. Juli 2007] . URL https://firstlook.org/theintercept/document/2015/07/01/unofficial-xks-user-guide/. Abruf: 30. November 2015.

[83] National Security Agency NSA. National Security Agency Technology Catalog. V3.0, März 2015. URL https://www.nsa.gov/research/_files/tech_transfers/nsa_technology_transfer_program.pdf. Abruf: 28. November 2015.

[84] National Security Agency NSA und Central Security Service CSS. Classification Guide for Human Language Technology (HLT) Models 2–20. *The Intercept*, 5. Mai 2015 [18. Mai 2011] . URL https://theintercept.com/document/2015/05/05/classification-guide-human-language-technology-hlt-models. Abruf: 27. November 2015.

[85] National Museum of Crime & Punishment. Forensic Linguistics & Author Identification. Identifying Someone's Personal Language. URL http://www.crimemuseum.org/crime-library/forensic-linguistics-and-author-identification/. Abruf: 25. November 2015.

[86] Der Spiegel (Online). Merkel zur Handy-Affäre: "Ausspähen unter Freunden - das geht gar nicht". 24. Oktober 2013. URL http://www.spiegel.de/politik/deutschland/handy-spaehaffaere-um-merkel-regierung-ueberprueft-alle-nsa-erklaerungen-a-929843.html. Abruf: 27. November 2015.

[87] Der Spiegel (Online). Auslandsgeheimdienst: BND spionierte Ministerien befreundeter Staaten aus. 7. November 2015. URL http://www.spiegel.de/politik/deutschland/bundesnachrichtendienst-spionierte-systematisch-freunde-aus-a-1061517.html. Abruf: 18. November 2015.

[88] Suraj Jung Pandey. D9.19 Report on methodology for applying existing machine learning methods for behavioural profiling. *Paper im Rahmen des 7. Europäischen Forschungrahmenprogramms FP7 (Projekt INDECT 218086)*, Juli 2011. URL http://www.indect-project.eu/files/deliverables/public/deliverable-9.19/at_download/file. Abruf: 27. November 2015.

[89] Suraj Jung Pandey. D9.9. Report on current state-of-the-art of machine learning methods for behavioural profiling. *Paper im Rahmen des 7. Europäischen Forschungrahmenprogramms FP7 (Projekt INDECT 218086)*, 29. April 2011. URL http://www.indect-project.eu/files/deliverables/public/deliverable-9.9/at_download/file. Abruf: 22. November 2015.

[90] Suraj Jung Pandey. D4.15 Framework for combining user supplied knowledge from diverse sources. *Paper im Rahmen des 7. Europäischen Forschungrahmenprogramms FP7 (Projekt INDECT 218086)*, Juli 2013. URL http://www.indect-project.eu/files/deliverables/public/deliverable-4.15/at_download/file. Abruf: 22. November 2015.

[91] Suraj Jung Pandey und Krzysztof Dorosz. D9.28 Methods for mining and detecting suspicious websites. *Paper im Rahmen des 7. Europäischen Forschungrahmenprogramms FP7 (Projekt INDECT 218086)*, Juli 2012. URL http://www.indect-project.eu/files/deliverables/public/deliverable-9.28/at_download/file. Abruf: 27. November 2015.

[92] Pavol Partila, Miroslav Vozňák, Adrián Kováč, und Michal Halas. Impact of Emotions on Fundamental Speech Signal Frequency. *Paper im Rahmen des 7. Europäischen Forschungrahmenprogramms FP7 (Projekt INDECT 218086)*, 2012. URL http://www.wseas.us/e-library/conferences/2012/Vienna/COMPUTERS/COMPUTERS-66.pdf. Abruf: 22. November 2015.

[93] Roland Peters. Blinde Spionage des BND: Was ist ein "Selektor"? *N-TV*, 20. Mai 2015. URL http://www.n-tv.de/politik/Was-ist-ein-Selektor-article15133656.html. Abruf: 30. November 2015.

[94] Kyle Rankin. NSA: Linux Journal is an "extremist forum" and its readers get flagged for extra surveillance. *Linux Journal*, 3. Juli 2014. URL http://www.linuxjournal.com/content/nsa-linux-journal-extremist-forum-and-its-readers-get-flagged-extra-surveillance. Abruf: 28. November 2015.

[95] Simon Rebiger. #NSAUA: Abgeordnete bekommen Einblick in BND-Selektoren. *Netzpolitik.org*, 17. November 2015. URL https://netzpolitik.org/2015/nsaua-abgeordnete-bekommen-einblick-in-bnd-selektoren/. Abruf: 18. November 2015.

[96] Marcel Rosenbach und Holger Stark. *Der NSA-Komplex: Edward Snowden und der Weg in die totale Überwachung*. Spiegel-Verlag, Hamburg, 2014. ISBN 978-3-641-14150-9.

[97] Anne Roth. Innenansicht einer Terrorismus-Ermittlung. *Das annalist Blog*. URL http://fallbeispiele.sozialebewegungen.org/annalist/. Abruf: 25. November 2015.

[98] Florian Rötzer. Auch die Schweiz hat ein Echelon-System. *Telepolis*, 25. November 1999. URL http://www.heise.de/tp/artikel/6/6532/1.html. Abruf: 28. November 2015.

[99] Joachim Scharloth. Überwachen und Sprache. How to do things with words. Video, 30. Chaos Communication Congress 30C3, 28. Dezember 2013. URL https://media.ccc.de/v/30C3_-_5377_-_de_-_saal_6_-_201312271245_-_uberwachen_und_sprache_-_josch#video. Abruf: 27. November 2015.

[100] Joachim Scharloth. 30c3 nachlese. *Surveillance and Security: Computer- und korpuslinguistische Methoden des politisch motivierten Internet-Monitorings*, 1. Januar 2014. URL http://www.security-informatics.de/blog/?p=1488. Abruf: 27. November 2015.

[101] Joachim Scharloth. 30C3 Nachlese, Teil 2. *Surveillance and Security: Computer- und korpuslinguistische Methoden des politisch motivierten Internet-Monitorings*, 8. Januar 2014. URL http://www.security-informatics.de/blog/?p=1499. Abruf: 27. November 2015.

[102] Joachim Scharloth. Die Geheimdienste lesen unsere E-Mails nicht! – Sie wissen aber trotzdem, was drin steht. *Surveillance and Security: Computer- und korpuslinguistische Methoden des politisch motivierten Internet-Monitorings*, 25. Mai 2014. URL http://www.security-informatics.de/blog/?p=1536. Abruf: 19. November 2015.

[103] Joachim Scharloth. Content Mapping mit Topic Models. *Surveillance and Security: Computer- und korpuslinguistische Methoden des politisch motivierten Internet-Monitorings*, 18. Februar 2015. URL http://www.security-informatics.de/blog/?p=1690. Abruf: 27. November 2015.

[104] Scharloth, Joachim. Gibt es einen sprachlichen Fingerabdruck? *Surveillance and Security: Computer- und korpuslinguistische Methoden des politisch motivierten Internet-Monitorings*, 21. September 2011. URL http://www.security-informatics.de/blog/?p=485. Abruf: 27. November 2015.

[105] Gerhard Schmid. Bericht über die Existenz eines Abhörsystems für private und wirtschaftliche Kommunikation (Abhörsystem ECHELON). 2011/2098 (INI). *Europäisches Parlament. Sitzungsdokument A5-0264/2001 Teil 1*, 11. Juli 2011. URL http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML+REPORT+A5-2001-0264+0+DOC+PDF+V0//DE. Abruf: 29. November 2015.

[106] Stefan Schmitt. Anonymisierung von Texten. Stilspuren verwischen. *Zeit Online*, 25. Juli 2013. URL http://www.zeit.de/2013/31/software-texte-anonymisierung. Abruf: 22. November 2015.

[107] Stefan Schulz. Chaos Communication Congress. Der radikalste Blogger im ganzen Land. *Frankfurter Allgemeine Zeitung FAZ (Online)*, 27. Dezember 2013. URL http://www.faz.net/aktuell/feuilleton/debatten/chaos-communication-congress-der-radikalste-blogger-im-ganzen-land-12728769.html#aufmacherBildJumpTarget. Abruf: 27. November 2015.

[108] Digitale Gesellschaft (Schweiz). Safe-Harbor-Urteil: Massenüberwachung verletzt den Wesensgehalt des Grundrechts auf Achtung des Privatlebens. 5. November 2015. URL https://www.digitale-gesellschaft.ch/2015/11/05/safe-harbor-urteil-massenueberwachung-verletzt-den-wesensgehalt-des-grundrechts-auf-achtung-des-privatlebens/. Abruf: 18. November 2015.

[109] Rote Hilfe Schweiz. Internationale Aktionstage: Marco libero! *Rote Hilfe Schweiz (Online)*, 31 Mai 2015. URL http://rotehilfech.noblogs.org/post/2015/05/31/internationale-aktionstage-marco-libero/. Abruf: 19. November 2015.

[110] Jürgen Seeger. Piraten im NRW-Landtag wollen Klarheit über "Zombie-Bügeleisen". *Magazin für professionelle Informationstechnik iX (Online)*, 10. Dezember 2013. URL http://www.heise.de/ix/meldung/Piraten-im-NRW-Landtag-wollen-Klarheit-ueber-Zombie-Buegeleisen-2063633.html. Abruf: 19. November 2015.

[111] Der Spiegel. NSA-Attacken auf SSL, VPN, SSH, Tor etc.: Das sind die Snowden-Dokumente. 29. Dezember 2014. URL http://www.spiegel.de/netzwelt/netzpolitik/snowden-dokumente-nsa-attacken-auf-ssl-vpn-ssh-tor-a-1010553.html. Abruf: 26. November 2015.

[112] StartPage. Unsere Datenschutzrichtlinien. URL https://startpage.com/deu/privacy-policy.html. Abruf: 26. November 2015.

[113] Jakob Steinschaden. Massenüberwachung zeigt soziale Folgen. 8. Mai 2015. URL https://www.freitag.de/autoren/netzpiloten/massenueberwachung-zeigt-soziale-folgen. Abruf: 27. November 2015.

[114] Martin Stoll. Schweizer Armee: Grosser Lauschangriff im All. *cryptome.org [SonntagsZeitung SZ]*, 9. Februar 1999. URL https://cryptome.wikileaks.org/jya/ch-wiretap.htm. Abruf: 28. November 2015.

[115] Susi Stühlinger. Wichtig zu Wissen: Ueli Maurers Kronjuwelen. Susi Stühlinger über allerlei österliche Vorkommnisse. *Wochenzeitung WOZ (Online)*, 9. April 2015. URL https://www.woz.ch/-5bf2. Abruf: 18. November 2015.

[116] Attila Szenogrady. «Rütli-Bomber»: Urbaniok bedroht – 21 Monate Haft. *20 Minuten 20min (Online)*, 25. Juni 2015. URL http://www.20min.ch/schweiz/zuerich/story/Urbaniok-bedroht---21-Monate-Haft-26909206. Abruf: 19. November 2015.

[117] Piotr Szwed, Wojciech Chmiel, Stanislaw Jędrusik, und Jacek Dańda. D6.4 ontology and automatic reasoning in crisis management – definitions and concepts. *Paper im Rahmen des 7. Europäischen Forschungrahmenprogramms FP7 (Projekt INDECT 218086)*, 28. Februar 2011. URL http://www.indect-project.eu/files/deliverables/public/d6.4/at_download/file. Abruf: 22. November 2015.

[118] Tageszeitung taz. Tatbestand Soziologie. Kommissar Google jagt Terroristen. 21. August 2007. URL http://www.taz.de/!5196250/. Abruf: 25. November 2015.

[119] Iain Thomson. Snowden scandal latest: NSA, GCHQ lingo-spies replaced by unstoppable RHINEHART robots. If you lost your job to software, you can sympathize. *The Register*, 6. Mai 2015. URL http://www.theregister.co.uk/2015/05/06/snowden_nsa_gchq_voice_transcription/. Abruf: 28. November 2015.

[120] Ueli Maurer. Ständerat – sommersession 2015 – achte sitzung – 11.06.15 – 08h15. 14.022. nachrichtendienstgesetz. zweirat. *Amtliches Bulletin – Die Wortprotokolle von Nationalrat und Ständerat*, 11. Juni 2015. URL http://www.parlament.ch/ab/frameset/d/s/4919/467625/d_s_4919_467625_467664.htm. Abruf: 29. November 2015.

[121] Erweiterte Universitätsleitung Universität Zürich. Leitbild der Universität Zürich. 29. November 2011. URL http://www.uzh.ch/dam/jcr:ffffffff-e1a6-549b-0000-00002247c2e3/uzh_leitbild_2012.pdf. Abruf: 27. November 2015.

[122] Institut für Computerlinguistik Universität Zürich. Was ist computerlinguistik? 12. September 2014. URL http://www.cl.uzh.ch/what-is-cl.html. Abruf: 27. November 2015.

[123] Pascal Unternährer. Marco Camenisch im offenen Vollzug. *Basler Zeitung BaZ (Online)*, 19. November 2015. URL http://bazonline.ch/zuerich/region/Marco-Camenisch-im-offenen-Vollzug/story/28679258. Abruf: 19. November 2015.

[124] Depesche "STATE 048489" USA. US embassy cables: Washington calls for intelligence on top UN officials. *The Guardian (Online)*, 28. November 2010. URL http://www.theguardian.com/world/us-embassy-cables-documents/219058. Abruf: 18. November 2015.

[125] WikiLeaks. Espionnage Élysée. Top French NSA Targets. 23. Juni 2015. URL https://wikileaks.org/nsa-france/selectors.html. Abruf: 23. November 2015.

[126] WikiLeaks. All the Chancellor's Men. NSA high priority targets for Germany. 20. Juli 2015. URL https://wikileaks.org/nsa-germany/selectors.html. Abruf: 22. November 2015.

[127] WikiLeaks. Target Tokyo. Top Japanese NSA Targets. 31. Juli 2015. URL https://wikileaks.org/nsa-germany/selectors.html. Abruf: 23. November 2015.

[128] Jaroslav Zdralek, Lukas Kapicak, Andrzej Figaj, Tadeusz Janasiewicz, Gema Maestro Molina, und Javier Sainz. Data Formats and Protocols for Information Handling in INDECT Portal. *Paper im Rahmen des 7. Europäischen Forschungrahmenprogramms FP7 (Projekt INDECT 218086)*, 7. Februar 2011. URL http://www.indect-project.eu/files/deliverables/public/deliverable-6.3/at_download/file. Abruf: 22. November 2015.

[129] Peng Zhong. PRISM Break Project. 12. Oktober 2015. URL https://prism-break.org. Abruf: 26. November 2015.



Philosophische Fakultät
Studiendekanat

Universität Zürich
Studiendekanat
Bereich Abschluss
Rämistr. 69
CH-8001 Zürich
Telefon +41 44 634 54 10
www.phil.uzh.ch

Selbstständigkeitserklärung

Hiermit erkläre ich, dass
die Masterarbeit von mir selbst und ohne unerlaubte Beihilfe verfasst worden ist und ich die Grundsätze wissenschaftlicher Redlichkeit einhalte (vgl. dazu: http://www.lehre.uzh.ch/index/LK-Plagiate-Merkblatt.pdf).

....

Ort und Datum Unterschrift

A Tabellen

A.1 Selektoren nach Modell

A.1.1 TFIDF

Aufbau	PNOS
aktion	2006
andi	pnos
frauen	rütli
prozess	schweizer
wef	sempach

Table A.1: Selektoren TFIDF-Modell: Einzelworte

Aufbau	PNOS
aktion flughafen	jonas gysin
antirassistische aktion	kanton bern
flughafen 20	kanton waadt
ibrahim abdallah	pnos ch
marco camenisch	tobias hirschi

Table A.2: Selektoren TFIDF-Modell: Wort-2-Gramme

Aufbau	PNOS
8 märz	1 august
abdallah ibrahim	8 märz
aktion flughafen	ausgaben rotem
camenisch marco	heinz kaiser
georges ibrahim	hirschi tobias

Table A.3: Selektoren TFIDF-Modell: 2–5-Wortkombinationen

A.1.2 Verdachtssprache

Aufbau	PNOS
eingestellter längst verfahren	meinungsfreiheit punkten stark
eingestellter längst wiederaufgreifen	möglichst nahe rechtsnorm
kampf mehrfach stark	möglichst nahe wortsinn
kriminalisiert mehrfach stark	möglichst rechtsnorm sinn
kriminalisiert repression stark	sicht spürbar wachstumswahns

Table A.4: Selektoren Verdachtssprache-Modell: intensivierend (generell)

Aufbau	PNOS
fluchthilfe leisten praktisch	aktionen positive sicherlich
fluchthilfe leisten praktisch unterstützen	mal öfters rein schauen
fluchthilfe praktisch unterstützen	mal rein schauen
frau nase voll	nationalen probleme rein
leisten praktisch unterstützen	öfters rein schauen

Table A.5: Selektoren Verdachtssprache-Modell: intensiverend (absolut)

Aufbau	PNOS
angegriffen kriminalisiert repression stark	allzu mehrheitsmeinung stark
angegriffen kriminalisiert stark	allzu stark systemeliten
angegriffen mehrfach stark	anwendung rechtsnormen stark
angegriffen repression stark	rechts sichtbar überprüfbar
kampf mehrfach stark	sicht spürbar wachstumswahns

Table A.6: Selektoren Verdachtssprache-Modell: intensivierend (hoch)

Aufbau	PNOS
derweil kampagnen möglichst	einheit kulturelle weitestgehend
flüchtlingsaufnahme kampagnen möglichst unattraktiv	einheit völkische weitestgehend
flüchtlingsaufnahme möglichst unattraktiv	extrem lebenslange verwahrung
kampagnen möglichst unattraktiv	festzusetzen möglichst nahe wortsinn
kampagnen möglichst versuchen	festzusetzen möglichst wortsinn

Table A.7: Selektoren Verdachtssprache-Modell: intensivierend (extrem hoch)

Aufbau	PNOS
aktionen militante vorwand	eigentlich unnütz verträge
angeklagten dienen eigentlich	führen gegner lügen
angeklagten dienen eigentlich entlastung	führen gegner lügen wahlkampf
angeklagten eigentlich entlastung	führen lügen wahlkampf
spiess tatsächlich umzudrehen	gegner lügen wahlkampf

Table A.8: Selektoren Verdachtssprache-Modell: Verschwörungsvokabular

Aufbau	PNOS
angegriffen grausam unzählige	261 nutzlos volkes
angegriffen grausam unzählige zivilistinnen	argumentation art inakzeptabel
angegriffen grausam zivilistinnen	argumentation inakzeptabel juristen
grausam niedermetzelten zivilistinnen	hand nehmen verfahren
grausam unzählige zivilistinnen	unsolidarisch verhält welt

Table A.9: Selektoren Verdachtssprache-Modell: skandalisierend

A.1.3 LDA

Aufbau	PNOS
wef stadt	pnos flüchtlingsflut
streik arbeiterinnen	leben bevölkerung
zürich frauen	schweizer schweiz
kapitalismus politik	stolperflüchtling stoppen
solidarität prozess	osama flüchtlingsflut

Table A.10: Selektoren LDA-Modell: 2-Wort-Topics

Aufbau	PNOS
arbeiterinnen demo aufbau	schweiz schweizer initiative
solidarität prozess politischen	flüchtlingsflut stoppen stolperflüchtling
zürich frauen märz	somit staat klar
stadt polizei basel	pnos flüchtlingsflut transparentaktion
leben politik kapital	leben politischen arbeit

Table A.11: Selektoren LDA-Modell: 3-Wort-Topics

Aufbau	PNOS
frauen zürich kapitalismus märz demo	ebenfalls politischen zudem sogar ziel
solidarität prozess gefangenen politischen revolutionären	pnos flüchtlingsflut lüthard august rütli
krieg regierung bewegung imperialistischen svp	schweiz schweizer steht usa grund
zürich streik arbeiterinnen aktion kampf	schweizer schweiz volk initiative somit
aufbau bern klar revolutionären steht	pnos flüchtlingsflut stoppen stolperflüchtling osama

Table A.12: Selektoren LDA-Modell: 5-Wort-Topics

Aufbau	PNOS
krieg regierung staat türkei politik imperialistischen bewegung partei	schweiz volk lassen staat politik politischen land parteien
streik arbeiterinnen kampf arbeiter unia streikenden klar bern	pnos flüchtlingsflut stolperflüchtling stoppen osama transparentaktion unterstützen partei
solidarität prozess gefangenen politischen revolutionären andi schweiz marco	schweizer schweiz initiative volkes svp usa armee franken
frauen kapitalismus märz kämpfen zürich gesellschaft aufbau krise	somit bevölkerung personen nationalen kinder art gesellschaft völker
zürich demo wef basel widerstand aktion stadt strasse	flüchtlingsflut transparentaktion osama stoppen pnos stolperflüchtling beitrag terrorist

Table A.13: Selektoren LDA-Modell: 8-Wort-Topics

Aufbau	PNOS
frauen märz kapitalismus arbeit kämpfen gesellschaft leben schweiz ausbeutung flugblatt	flüchtlingsflut pnos osama transparentaktion stolperflüchtling stoppen unterstützen terrorist beitrag aktivisten
solidarität zürich streik schweiz andi aufbau gefangenen revolutionären prozess bellinzona	schweiz steht lassen arbeit staat usa land politischen europa gesellschaft
zürich demo repression widerstand basel strasse polizei raum demonstration winterthur	initiative somit svp zudem klar franken souverän kinder volksinitiative schutz
krieg marco kampf regierung revolution türkei imperialistischen knast rojava gefangenen	pnos partei kanton bern lüthard august rütli stoppen langenthal medien
arbeiterinnen wef arbeiter politik leute unia krise interessen klar steht	schweizer volk schweiz politik volkes leben grund armee meinung politische

Table A.14: Selektoren LDA-Modell: 10-Wort-Topics

B Abstracts computerlinguistisch relevanter NSA-Patente

Nachfolgend sind einige US-Patente in ihren Abstacts ausgeführt, die von der NSA angemeldet wurden und die aufzeigen, dass diese schon seit Jahrzehnten computerlinguistische oder (rein) statistische Verfahren erforscht und schützt, die im Zusammenhang mit der Verarbeitung natürlichsprachlicher Daten stehen.

Alle im Volltext bei *Google* einsehbaren Patente wurden am 21. November 2015 abgerufen.

B.1 Method of retrieving documents that concern the same topic (1995)

A method of identifying, retrieving, or sorting documents by language or topic involving the steps of creating an n-gram array for each document in a database, parsing an unidentified document or query into n-grams, assigning a weight to each n-gram, removing the commonality from the n-grams, comparing each unidentified document or query to each database document, scoring the unidentified document or query against each database document for similarity, and based on the similarity score, identifying retrieving, or sorting the document or query with-respect to language or topic.

Volltext unter URL `https://www.google.com/patents/US5418951`

B.2 Language-independent method of generating index terms (1998)

Index terms are drawn from text documents without the need for language-specific processes or training and are suitable as gists for the subject documents. Index terms are extracted on the basis of scores of constituent n-grams relative to n-gram counts in a corpus. A method of extracting joint index terms to represent a plurality of documents is also provided.

Volltext unter URL `https://www.google.com/patents/US5752051`

B.3 Automatically generating a topic description for text and searching and sorting text by topic using the same (1999)

A method of automatically generating a topical description of text by receiving the text containing input words; stemming each input word to its root form; assigning a user-definable part-of-speech score to each input word; assigning a language salience score to each input word; assigning an input-word score to each input word; creating a tree structure under each input word, where each tree structure contains the definition of the corresponding input word; assigning a definition-word score to each definition word; collapsing each tree structure to a corresponding tree-word list; assigning a tree-word-list score to each entry in each tree-word list; combining the tree-word lists into a final word list; assigning each word in the final word list a final-word-list score; and choosing the top N scoring words in the final word list as the topic description of the input text. Document searching and sorting may be accomplished by performing the method described above on each document in a database and then comparing the similarity of the resulting topical descriptions.

Volltext unter URL `https://www.google.com/patents/US5937422`.

B.4 Device and method for full-text large-dictionary string matching using n-gram hashing (2001)

A method and apparatus providing full-text scanning for matches in a large dictionary is described. The invention is suitable for SDI (selective dissemination of information) systems, accommodating large dictionaries (104 to 105 entries) and rapid processing. A preferred embodiment employs a hardware primary test on a single commercially-available gate-array board hosted by a computer, in which a software secondary test is conducted. No delimiter cues such as spaces or punctuation are required.

Volltext unter URL `https://www.google.com/patents/US6169969`.

B.5 Method for finding large numbers of keywords in continuous text streams (2001)

A method of full-text scanning for matches in a large dictionary of keywords is described, suitable for SDI (selective dissemination of information). The method is applicable to large dictionaries (hundreds of thousands of entries) and to arbitrary byte sequences for both patterns and sample streams. The approach employs Boyer-Moore-Horspool skipping, extended to pattern collections and digrams, followed by an n-gram hash test, which also identifies a subset of feasible keywords for conventional pattern matching at each location of a putative

match.

Volltext unter URL https://www.google.com/patents/US6311183.

B.6 Method of summarizing text using just the text (2005)

A method of summarizing a text by the following steps. Identifying the textual units in the text. Selecting a first set of textual units and identifying its textual units. Selecting a second set of textual units and identifying its textual units. Determining how many textual units are shared between the first and second sets of textual units. Selecting a third set of textual units between the first and second set of textual units and identifying its unique textual units. Determining the frequency of occurrence of the textual unit in the third set of textual units. Determining the frequency of occurrence of the textual unit in the text. Determining the proximity of the results of the last two steps. Calculating a score for the first set of textual units. Assigning the highest score to the first set of textual units. Selecting a numbers of first sets of textual units, according to score, as the summary of the text.

Volltext unter URL https://www.google.com.ar/patents/US6904564.

B.7 Method of summarizing text by sentence extraction (2006)

A method of summarizing text. The sentences in the text are identified first. Then, the terms in each sentence are identified. A matrix is then generated, where the columns represent the sentences and the rows represent the terms. The entries in the matrix are weighted with an exponentially decaying function or a Hidden Markov Model. The Euclidean length of each column is determined. The sentence corresponding to the column having the maximum Euclidean length is selected as a summary sentence. The columns corresponding to the remaining sentences have their matrix entries reduced. If additional summary sentences are desired then return to the step of determining Euclidean length of the columns.

Volltext unter URL https://www.google.com/patents/US6990634.

B.8 Method of optical character recognition using feature recognition and baseline estimation (2008)

The present invention is a method of optical character recognition. First, text is received. Next all words in the text are identified and associated with the appropriate line in the document. The directional derivative of the pixellation density function defining the text is then taken,

and the highest value points for each word are identified from this equation. These highest value points are used to calculate a baseline for each word. A median anticipated baseline is also calculated and used to verify each baseline, which is corrected as necessary. Each word is then parsed into feature regions, and the features are identified through a series of complex analyses. After identifying the main features, outlying ornaments are identified and associated with appropriate features. The results are then compared to a database to identify the features and then displayed.

Volltext unter URL https://www.google.com/patents/US7454063.

B.9 Natural language database searching using morphological query term expansion (2010)

The present invention is a method of database searching. First, a language is selected. Next, elements are received. The system is then searched to identify at least one unit number that is associated with a chosen element, the unit number being linked to a data unit containing morphological variants of the element. If no unit number is identified, the element is compared to a prefix list. If no match is found there, the element is broken into a prefix and suffix, and the prefix and suffix are matched to a prefix list, suffix list or a unit number. This process is repeated for all elements. A unit number associated with each element is then chosen, and the elements contained in the data units linked to the unit numbers are compared to a database. The results are displayed and preferably ranked according to user preferences. If an element is associated with multiple unit numbers, this process is repeated until all data units have been compared to the database.

Volltext unter URL https://www.google.com/patents/US7761286.

B.10 Method of database searching (2010)

The present invention is a method of database searching. First, a language is selected and elements received. The system is searched to identify a unit number associated with each element, which is linked to a data unit containing morphological variants of the element. If none are identified, the element is broken into sub-textual units that may contain a prefix, compound-prefix, and/or suffix along with a primary element. A unit number is then obtained for the primary element. If this does not result in a match, the elements may be saved in a database for further linguistic development. A unit number associated with each matched element is then chosen, and the elements contained in the data units linked to the unit numbers are compared to a database index. If an element is associated with multiple unit numbers, this process is repeated until all data units have been compared to the database.

Volltext unter URL https://www.google.com/patents/US7797152.

B.11 Method of identifying topic of text using nouns (2010)

A method of identifying a topic of a text. Text is received. Then, the nouns in the text are identified. The singular form of each identified noun is determined. Combinations are created of the singular form of the identified nouns, where the number of singular forms of the nouns in the combinations is user-definable. The frequency of occurrence in the text of each noun that corresponds to its singular form is determined. Each frequency of occurrence is assigned as a score to its corresponding singular form noun. Each combination of singular form nouns is assigned a score that is equal to the sum of the scores of its constituent singular form nouns. The user-definable number of top scoring singular form nouns and combinations of singular form nouns are selected as the topic of the text.

Volltext unter URL https://www.google.com/patents/US7805291.

B.12 Method of assessing language translation and interpretation (2012)

A method of assessing quality of language translation and interpretation by receiving source material and a translation, identifying the source material's content and format, assigning a first rating to the source material's level of difficulty in translating the source material, determining the translation's type, assigning a second rating to the translation's accuracy, assigning a third rating to the degree to which the translation interprets the source material's intended message, assigning a fourth rating to the formatting of the translation, and evaluating the four ratings to determine an assessment of the translation's language translation and interpretation.

Volltext unter URL https://www.google.com/patents/US8185373.

B.13 Device for and method of language processing (2013)

The present invention is a device for and method of language processing that includes a communication database of communications, a transcription database of transcripts for the communication, an extractor for extracting a visual representation of each communication, a first displayer for displaying a visual representation of a communication and its transcription, a segmentor for segmenting a visual representation, a media player, a first editor for

blanking portions of a transcription and adding text, a second editor for filling in blanks and adding text, a second displayer for displaying a transcription that were blanked along with the corresponding entries made by the second editor and adding textual information, and a third displayer for providing feedback.

Volltext unter URL `https://www.google.com/patents/US8380485`.

C Inoffizielles oder unpubliziertes Quellenmaterial

C.1 XKeyscore-Regeln im Zusammenhang mit Tor-Anonymisierungstechnologien

Am 29. November 2015 unter http://daserste.ndr.de/panorama/xkeyscorerules100.txt abgerufen:

```
// START_DEFINITION
/**
 * Fingerprint Tor authoritative directories enacting the directory protocol.
 */
fingerprint('anonymizer/tor/node/authority') = $tor_authority
  and ($tor_directory or preappid(/anonymizer\/tor\/directory/));
// END_DEFINITION

// START_DEFINITION
/*
Global Variable for Tor foreign directory servers. Searching for potential Tor
clients connecting to the Tor foreign directory servers on ports 80 and 443.
*/

$tor_foreign_directory_ip = ip('193.23.244.244' or '194.109.206.212' or
'86.59.21.38' or '213.115.239.118' or '212.112.245.170') and port ('80' or
'443');
// END_DEFINITION

// START_DEFINITION
/*
this variable contains the 3 Tor directory servers hosted in FVEY countries.
Please do not update this variable with non-FVEY IPs. These are held in a
separate variable called $tor_foreign_directory_ip. Goal is to find potential
Tor clients connecting to the Tor directory servers.
*/
$tor_fvey_directory_ip = ip('128.31.0.39' or '216.224.124.114' or
'208.83.223.34') and port ('80' or '443');
// END_DEFINITION


// START_DEFINITION
requires grammar version 5
```

```
/**
 * Identify clients accessing Tor bridge information.
 */
fingerprint('anonymizer/tor/bridge/tls') =
ssl_x509_subject('bridges.torproject.org') or
ssl_dns_name('bridges.torproject.org');

/**
 * Database Tor bridge information extracted from confirmation emails.
 */
fingerprint('anonymizer/tor/bridge/email') =
email_address('bridges@torproject.org')
  and email_body('https://bridges.torproject.org/' : c++
  extractors: {{
    bridges[] = /bridge\s([0-9]{1,3}\.[0-9]{1,3}\.[0-9]{1,3}\.[0-9]{1,3})
        :?([0-9]{2,4}?[^0-9])/;
  }}
  init: {{
    xks::undefine_name("anonymizer/tor/torbridges/emailconfirmation");
  }}
  main: {{
    static const std::string SCHEMA_OLD = "tor_bridges";
    static const std::string SCHEMA_NEW = "tor_routers";
    static const std::string FLAGS = "Bridge";
    if (bridges) {
      for (size_t i=0; i < bridges.size(); ++i) {
        std::string address = bridges[i][0] + ":" + bridges[i][1];
        DB[SCHEMA_OLD]["tor_bridge"] = address;
        DB.apply();
        DB[SCHEMA_NEW]["tor_ip"] = bridges[i][0];
        DB[SCHEMA_NEW]["tor_port_or"] = bridges[i][1];
        DB[SCHEMA_NEW]["tor_flags"] = FLAGS;
        DB.apply();
      }
      xks::fire_fingerprint("anonymizer/tor/directory/bridge");
    }
    return true;
  }});
// END_DEFINITION


// START_DEFINITION
/*
The fingerprint identifies sessions visiting the Tor Project website from
non-fvey countries.
*/
fingerprint('anonymizer/tor/torpoject_visit')=http_host('www.torproject.org')
and not(xff_cc('US' OR 'GB' OR 'CA' OR 'AU' OR 'NZ'));
// END_DEFINITION


// START_DEFINITION
```

```
/*
These variables define terms and websites relating to the TAILs (The Amnesic
Incognito Live System) software program, a comsec mechanism advocated by
extremists on extremist forums.
*/

$TAILS_terms=word(’tails’ or ’Amnesiac Incognito Live System’) and word(’linux’
or ’ USB ’ or ’ CD ’ or ’secure desktop’ or ’ IRC ’ or ’truecrypt’ or ’ tor ’);
$TAILS_websites=(’tails.boum.org/’) or (’linuxjournal.com/content/linux*’);
// END_DEFINITION

// START_DEFINITION
/*
This fingerprint identifies users searching for the TAILs (The Amnesic
Incognito Live System) software program, viewing documents relating to TAILs,
or viewing websites that detail TAILs.
*/
fingerprint(’ct_mo/TAILS’)=
fingerprint(’documents/comsec/tails_doc’) or web_search($TAILS_terms) or
url($TAILS_websites) or html_title($TAILS_websites);
// END_DEFINITION


// START_DEFINITION
requires grammar version 5
/**
 * Aggregate Tor hidden service addresses seen in raw traffic.
 */
mapreduce::plugin(’anonymizer/tor/plugin/onion’) =
  immediate_keyword(/(?:([a-z]+):\/\/){0,1}([a-z2-7]{16})\.onion(?::(\d+)){0,1}/c :
      c++
    includes: {{
      #include <boost/lexical_cast.hpp>
    }}
    proto: {{
      message onion_t {
        required string address = 1;
        optional string scheme = 2;
        optional string port = 3;
      }
    }}
    mapper<onion_t>: {{
      static const std::string prefix = "anonymizer/tor/hiddenservice/address/";

      onion_t onion;
      size_t matches = cur_args()->matches.size();
      for (size_t pos=0; pos < matches; ++pos) {
        const std::string &value = match(pos);
        if (value.size() == 16)
          onion.set_address(value);
        else if(!onion.has_scheme())
          onion.set_scheme(value);
```

```
        else
          onion.set_port(value);
      }

      if (!onion.has_address())
        return false;

      MAPPER.map(onion.address(), onion);
      xks::fire_fingerprint(prefix + onion.address());
      return true;
    }}
    reducer<onion_t>: {{
      for (values_t::const_iterator iter = VALUES.begin();
          iter != VALUES.end();
          ++iter) {
        DB["tor_onion_survey"]["onion_address"] = iter->address() + ".onion";
        if (iter->has_scheme())
          DB["tor_onion_survey"]["onion_scheme"] = iter->scheme();
        if (iter->has_port())
          DB["tor_onion_survey"]["onion_port"] = iter->port();
        DB["tor_onion_survey"]["onion_count"] = boost::lexical_cast<std::string>(
            TOTAL_VALUE_COUNT);
        DB.apply();
        DB.clear();
      }
      return true;
    }});

/**
 * Placeholder fingerprint for Tor hidden service addresses.
 * Real fingerpritns will be fired by the plugins
 *   'anonymizer/tor/plugin/onion/*'
 */
fingerprint('anonymizer/tor/hiddenservice/address') = nil;
// END_DEFINITION


// START_DEFINITION
appid('anonymizer/mailer/mixminion', 3.0, viewer=$ascii_viewer) =
        http_host('mixminion') or
        ip('128.31.0.34');
// END_DEFINITION
```

C.2 Daniel Mossbrucker (2015): Digitale Informantenschutzrechte

Deckblatt, Zusammenfassung, Inhaltsverzeichns und Fazit

Nachfolgend werden mit ausdrücklicher Genehmigung von Daniel Mossbrucker folgende Teile seiner Bachelorarbeit zur Verfügung gestellt:

- Deckblatt nach den Vorgaben der Technischen Universität Dortmund
- Zusammenfassung (Abstract)
- Volles Inhaltsverzeichnis
- Fazit
- Literaturverzeichnis

Auf Grund des Sperrvermerks zum Schutz besonders schützenswerter persönlicher Informantendaten, die unter dem journalistischen Quellenschutz fallen, kann die Arbeit in ihrem Hauptteil zur Zeit nicht und später wenn – dann nur eingeschränkt – publiziert werden.

Nichtsdestotrotz können im Anhang die interessierenden Quellenverweise, die wichtigen Befunde (summarisch) und damit Schlüsse der Arbeit öffentlich gemacht werden.

Für die Möglichkeit diese wichtige Arbeit in Auszügen in den Anhang stellen zu können, danke ich Daniel Mossbrucker an dieser Stelle ausdrücklich.

Technische Universität Dortmund
Fakultät für Kulturwissenschaften
Institut für Journalistik

Digitale Informantenschutzrechte

Was offenbaren hinterlassene digitale Daten über die Umstände einer journalistischen Recherche und wie sind Journalisten vor einem Zugriff auf diese Daten durch Ermittlungsbehörden rechtlich geschützt?

Bachelorarbeit
im Studiengang
Journalistik

Betreuer/in: Prof. Dr. Tobias Gostomzyk
Zweitprüfer/in: Prof. Dr. Frank Lobigs
vorgelegt von: Daniel Moßbrucker
vorgelegt am: 30.07.2015

Zusammenfassung

Die vorliegende Arbeit behandelt die Frage, inwiefern sich durch die Digitalisierung die journalistische Recherche so verändern hat, dass die rechtlichen Voraussetzungen in Deutschland nicht mehr ausreichen, um einen umfassenden publizistischen Informantenschutz zu gewährleisten. Mobilfunk und Internet haben Journalisten Recherchemöglichkeiten beschert, die im analogen Zeitalter undenkbar schienen. Die dabei hinterlassenen Kommunikationsdaten ermöglichen es jedoch Geheimdiensten, Polizeibehörden und Strafermittlern, den Verlauf von journalistischen Recherchen exakt nachzuzeichnen und damit Kommunikationsverhältnisse mit Informanten offenzulegen.

Damit stehen Ermittlern Maßnahmen mit bisher ungekannten Möglichkeiten zur Verfügung. Durch die Normen, die klassischerweise den Informantenschutz in Deutschland begründen, sind Journalisten davor nicht geschützt. Das Zeugnisverweigerungsrecht für Medienschaffende (§ 53 StPO) sowie ein Durchsuchungs- und Beschlagnahmeverbot von Redaktionsräumen (§ 97 StPO) beschränkt sich auf die analoge Welt. Auf Kommunikationsdaten der Mobilfunk- und Internetnutzung von Journalisten kann hingegen einfacher zugegriffen werden, da in der Gesetzgebung hier eine bewusste Abstufung zwischen Journalisten und anderen Berufsgeheimnisträgern vollzogen worden ist (§ 160a StPO)

Dieses in der Literatur kritisierte Defizit wird in einem Feldexperiment in die Realität überführt, indem die Recherche eines Journalisten technisch überwacht wird. Die anfallenden Kommunikationsdaten werden anschließenden vor dem Hintergrund ausgewertet, was sie über den Journalisten und seine Beziehung zu seinen Informanten aussagen. Dabei wird deutlich, dass bereits die Verkehrsdaten ausreichen, um ein Kommunikations- und Bewegungsprofil des Journalisten zu skizzieren. Dieses dürfte in einem Strafverfahren ausreichen, um weitergehende Ermittlungsmaßnahme wie die Auswertung von Kommunikationsinhalten zu genehmigen und damit Rechtsverstöße von Informanten nachzuweisen.

Die Untersuchung kommt zu dem Ergebnis, dass im Falle des Feldexperiments die rechtlichen Normen der Strafprozessordnung ausgereicht haben dürften, um Informanten zu schützen. Dies liegt daran, dass die von den Informanten begangenen Straftaten nicht im Katalog der Straftaten aufgeführt werden, die eine Verkehrsdatenabfrage erlauben. Für andere Recherchen kann nicht ausgeschlossen werden, dass Journalisten als Teilnehmer einer solchen Straftat eingestuft werden und sie Ziel von Ermittlungsmaßnahmen werden.

Die Ergebnisse verdeutlichen, dass in Vorhaben wie der Vorratsdatenspeicherung eine Gefährdung für die journalistische Arbeit zu sehen ist. Gleichzeitig muss Informantenschutz im Zeitalter der Digitalisierung neu gedacht und durch verantwortungsvolles Handeln der Journalisten – etwa in Form von verschlüsselter Kommunikation – umgesetzt werden.

Inhaltsverzeichnis

Dank I

Sperrvermerk II

Gleichstellungsvermerk II

Hinweise zur verwendeten Literatur II

Zusammenfassung III

Abbildungsverzeichnis V

Tabellenverzeichnis VI

Abkürzungsverzeichnis VII

1 Einleitung 1

2 Ausgangslage und Problemstellung 3

2.1 Soziales Problem 3

2.2 Rechtliche Ausgangslage: Grundlagen des Informantenschutzes in Deutschland 3

2.2.1 Absicherung des Informantenschutzes durch das publizistische Zeugnisverweigerungsrecht sowie Durchsuchungs- und Beschlagnahmeverbote 4

2.2.1.1 Zeugnisverweigerungsrecht 4

2.2.1.2 Beschlagnahmeverbot 6

2.2.2 Historische Entwicklung und inhaltliche Begründung publizistischer Informantenschutzrechte durch die Rechtsprechung des Bundesverfassungsgerichts 7

2.2.4 Schutz der Presse nach Art. 5 9

2.2.5 Fernmeldegeheimnis nach Art. 10 GG und die Anwendbarkeit auf publizistische Informantenschutzrechte nach §§ 53 Abs. 1 Satz 1 Nr. 5 StPO und 97 Abs. 5 StPO 10

2.2.6 Zwischenfazit: publizistische Informantenschutzrechte durch prozessuale Normen der §§ 53, 97 StPO sowie der Art. 5, 10 GG 11

3 Entwicklung und Bedeutungsgewinn des Internets 13

3.1 Historische Entwicklung und Bedeutungsgewinn des Internets für Gesellschaft und Journalismus 13

3.2 Die Technik des Internets 17

3.2.1 Internet Protocol und IP-Adresse 17

3.2.2 Domain Name System 20

3.2.3 Routing 20

3.2.4 E-Mail 21

3.2.7 Zwischenbetrachtung: Die Relevanz des Internets für die journalistische Recherche und damit verbundene Datenspuren 23

4 Auswirkungen verdeckter Ermittlungsmaßnahmen auf Informantenschutzrechte 24

4.1 Begriffsbestimmung: Telekommunikation, Verkehrsdaten, Bestandsdaten 24

IV

Inhaltsverzeichnis

4.2 Normen der StPO zu verdeckten Ermittlungsmaßnahmen 27
4.2.1 Überwachung der Telekommunikation nach § 100a StPO 27
4.2.2 Erhebung von Telekommunikationsverbindungsdaten nach § 100g StPO..................30
4.2.3 Bestandsdatenauskunft nach § 100j StPO 31
4.2.4 Weitere Eingriffe in das Fernmeldegeheimnis gemäß §§ 100f, 100h, 100i StPO und die Bedeutung für den Journalismus 32
4.3 Exkurs: Das Urteil des Bundesverfassungsgerichts zur Vorratsdatenspeicherung.....34
4.4 Zwischenbetrachtung: Auswirkungen auf publizistische Informantenschutzrechte durch strafprozessuale Normen der verdeckten Ermittlung 36
4.5 Präventive Maßnahmen in den Polizei- und BKA-Gesetzen..37
4.6 Grundlagen publizistischer Informantenschutzrechte 38
5 Empirische Fallstudie zur Bedeutung digitaler Daten für Informantenschutzrechte 40
5.1 Forschungsfrage 41
5.2 Operationalisierung des Forschungsinteresses..42
5.2.1 Leitfragen für die Untersuchung 42
5.2.2 Begriffsbestimmungen 43
5.2.2.1 Telekommunikation, Verkehrsdaten und Bestandsdaten 43
5.2.2.2 IP, IP-Adresse, DNS, Routing, Logfile, Cookies, E-Mail (rechtlich und technisch).......... 43
5.2.2.3 Daten 43
5.2.3 Bildung von Indikatoren 44
5.2.4 Wahl der wissenschaftlichen Methode 46
5.2.5. Entwicklung des Instruments...49
5.2.6 Pretest und Instrumentenprüfung...53
5.2.7 Stichprobenziehung...53
5.3 Datenerhebung...54
5.5 Datenauswertung...56
5.6 Ergebnisse der Untersuchung 57
5.6.1 Telekommunikationsverbindungsdaten im Sinne von Telefonie (Verkehrsdaten).......58
5.6.2 Telekommunikationsverbindungsdaten im Sinne von Kurzmitteilungen (Verkehrsdaten)...62
5.6.3 Ergebnisse: Telekommunikationsverbindungsdaten im Sinne von Kurzmitteilungen (Verkehrsdaten)...63
5.6.3.1 Zwischenbetrachtung: Kommunikationsverhalten mit Informanten über Telefon, Kurzmitteilung und E-Mail 64
5.6.4 Verwendete IP-Adressen während der Recherche...65
5.6.5 Netzwerkpaketdaten der Internetnutzung 66
5.6.4 besuchte Websites 70
5.6.5 Bestandsdaten..72
5.7 Interpretation der Ergebnisse 72
5.7.1 Aussagekraft der gesammelten Kommunikationsdaten 73
5.7.1.1 Rückbezug auf die Leitfragen bezüglich der Aussagekraft digitaler Daten........................ 75
5.7.2 Rechtliche Bewertung des Feldexperimentes 77

IV

Inhaltsverzeichnis

5.7.2.1 Rückbezug auf die Leitfrage bezüglich der Anwendbarkeit der §§ 100a, 100g und 100j StPO 78
5.7.3 Interpretation der Ergebnisse vor dem Hintergrund der Forschungsfrage 79

6 Methodenkritik 82

7 Fazit und Ausblick 84

Anhang A: Ergänzende Darstellungen zur Technik des Internets (Logfiles und Cookies) 87

A.1 Log-Dateien 87

A.2 Cookies 88

Anhang B: Konzeption der Inhaltsanalyse zur Datenauswertung 91

B.1 Allgemeine Festlegungen zur Daten- und Analyseart 91

B.2 Herleitung der Kategoriensysteme 95

B.2.1 Kategoriensystem für Telekommunikationsverbindungsdaten im Sinne von Telefonie (Verkehrsdaten) 96
B.2.2 Kategoriensystem für Telekommunikationsverbindungsdaten im Sinne von Kurznachrichten (Verkehrsdaten) 97
B.2.3 Kategoriensystem für Telekommunikationsverbindungsdaten im Sinne von E-Mail (Verkehrsdaten) 98
B.2.4 Kategoriensystem für Internetnutzungsdaten im Sinne von besuchten Webseiten 99
B.2.5 Kategoriensystem für Cookies 100
B.2.6 Kategoriensystem für IP-Adressen 101
B.2.7 Kategoriensystem für Netzwerkdaten 102

Anhang C: Paraphrasierendes Zusammenfassen „besuchte Webseiten“ 106

Anhang D: Codebuch für die Datenauswertung 111

Anhang E: Ergänzende Ergebnisse und Datensätze 114

Literaturverzeichnis 119

Eidesstattliche Versicherung 123

7 Fazit und Ausblick

Diese Arbeit warf zu Beginn die ganz allgemeine Frage auf, ob Journalisten heute noch ihren Informanten glaubwürdig Quellenschutz zusichern können. Die Spiegel-Affäre ist über ein halbes Jahrhundert her, seitdem hat das Bundesverfassungsgericht in ständiger Rechtsprechung den publizistischen Informantenschutz und damit die Pressefreiheit gestärkt. Deutschland rangiert weltweit auf Platz zwölf der Pressefreiheit[211] und eine kollektive Angst der Journalisten vor Überwachung ist trotz NSA-Affäre nicht auszumachen. Ob diese Haltung noch zeitgemäß ist, muss nach der vorliegenden Untersuchung ernsthaft bezweifelt werden.

Im Wesentlichen waren mit der Arbeit zwei Erkenntnisziele verbunden: Es sollte erstens geklärt werden, was digitale Daten über die journalistische Recherche aussagen können und zweitens, ob rechtlich die Zugriffsbefugnisse von Strafermittlern auf diese Kommunikationsdaten den Informantenschutz untergraben können. Allen in der Methodenkritik aufgeführten Einschränkungen vorangestellt muss festgehalten werden, dass sowohl die Kommunikationsdaten an sich wie auch der Umgang mit ihnen das Potential haben, publizistische Informantenschutzrechte zu schwächen.

Im rechtlichen Bereich konnte gezeigt werden, dass es hier theoretisch zu einer Schwächung der Informantenschutzrechte gekommen ist. Dass Journalisten im analogen Bereich quasi absolute Schutzrechte durch das Zeugnisverweigerungsrecht und das Beschlagnahmeverbot genießen, dieser Schutz aber bei verdeckten Ermittlungsmaßnahmen abgeschwächt worden ist, wird in der Literatur als unverständlich kritisiert. Erst Recht wenn man sieht, dass andere Berufsgeheimnisträger wie Ärzte, Rechtsanwälte und Geistliche ihre Schutzrechte auch im digitalen Bereich in gleicher Qualität behalten haben. Von einer grundsätzlichen Bedrohung des journalistischen Quellenschutzes zu sprechen, mag sicher übertrieben sein. Aber dass es schon rein quantitativ mehr Möglichkeiten gibt, mit verdeckten Methoden gegen Journalisten zu ermitteln als mit offenen, ist offensichtlich und im Sinne der Pressefreiheit sicherlich ein Rückschritt.

Diese Entwicklung ist umso bedrohlicher, wenn man sich anschaut, was die Datenspuren des Journalisten im vorliegenden Fall über seine Recherche offenbart haben. Beziehungsgeflechte, Bewegungsprofile und Verhaltensmuster – digitale Daten ermöglichen es, einer Recherche alle Vertraulichkeit zu nehmen. Dieses Potential im Einzelfall sichtbar gemacht zu haben, ist der zentrale Erkenntnisgewinn dieser Arbeit und kann helfen, einer Debatte, die bisher vor allem theoretisch im juristischen Fachdiskurs geführt worden ist, ein greifbares und anschauliches Exemplum zu liefern.

Die Ergebnisse sollten Ermutigung genug sein, sich dieses für den Journalismus und die Gesellschaft insgesamt wichtigen Themas weiter wissenschaftlich zu widmen. Aufgrund

[211] vgl. Reporter ohne Grenzen (2015): 1

der formalen, finanziellen und technischen Beschränkungen dieser Arbeit unterliegen die Ergebnisse beschränkter Aussagekraft, die es in weiteren Arbeiten zu fundieren gilt. Hier hat die Arbeit vielschichtige Möglichkeiten offenbart: Zunächst steckt in den Kommunikationsdaten selbst ein kaum greifbares Potential, das hier nur zu Bruchteilen ausgewertet werden konnte. Speziell die über 39 Millionen Netzwerkpaketdaten bieten die Möglichkeit, Informanten-Beziehungen, die über das Internet geführt worden sind, noch zielgerichteter zu analysieren. Mit einer Deep Packet Inspection (DPI) könnte beispielsweise gezeigt werden, was alles aus den Paketdaten zu lesen ist[212] – und möglicherweise auch schon gelesen wird[213]. Bei wiederholenden Experimenten wäre es angebracht, einerseits die Zahl der Probanden zu erhöhen und andererseits die technische Ausstattung, insbesondere von Mobilfunkgeräten, zu verbessern, um den Umfang der Datensammlung den realen Bedingungen weiter anzugleichen. Beides zusammengenommen würde die Aussagekraft der Experimente weiter stärken.

Im qualitativen Bereich ist sichtbar geworden, dass eine Beschränkung auf Normen der StPO zwar aus historischen Gesichtspunkten sinnvoll ist, gleichzeitig jedoch polizeilich-präventive Maßnahmen sowie die Möglichkeiten der Geheimdienste dabei ausgeklammert werden. Im Lichte der NSA-Veröffentlichungen und des Falls Vorbeck beim Nachrichtenmagazin „Der Spiegel“[214] wird offensichtlich, wie notwendig eine Ausdehnung von Untersuchungen auf diese Gesetze wäre.

Da das Bedrohungspotential, das in digitalen Daten für den publizistischen Informantenschutz stecken kann, immer offensichtlicher wird, wäre es überdies ebenfalls lohnend, den Blick auf die Akteure zu richten: Fühlen sich investigative Journalisten durch mögliche Überwachungsmaßnahmen in ihrer Arbeit eingeschränkt? Eine qualitative Befragung könnte zeigen, ob die zumeist abstrakte Gefahr schon dazu geführt, dass die Pressefreiheit auf diese Weise eingeschränkt worden ist.

Auf die Digitalisierung mit einem solchen vorauseilendem Gehorsam zu antworten, wäre zweifelsfrei die falsche Reaktion. Gleichzeitig ist aber natürlich zu beachten, dass sich die Entwicklung nicht zurückdrehen lassen wird. Es wäre aus journalistischer Perspektive auch gar nicht wünschenswert. Die Verdatung der Welt, das Internet der Dinge und die zunehmende Bedeutung von Query-Öffentlichkeit sind Trends, die sich eher beschleunigen werden als zurückgehen. Die Vorteile, die darin liegen, werden von der Bevölkerung (noch) als zu wertvoll angesehen, um sich wegen der Gefahren solchen technologischen Neuerungen zu widersetzen.

Ein passendes Beispiel vor dem Hintergrund dieser Arbeit ist sicherlich die Debatte um eine Wiedereinführung der Vorratsdatenspeicherung, die bei Fertigstellung dieser Arbeit

[212] Für eine umfassende technische und rechtliche Einführung in die DPI vgl. Bedner (2009).

[213] vgl. Meister (2012).

[214] vgl. „Der Spiegel“, „Anschlag auf die Pressefreiheit“, Heft 28/2015.

intensiv geführt wird.[215] Die Ergebnisse des Feldexperiments haben gezeigt, dass im vorliegenden Fall die Kommunikationsbeziehung zwischen Journalist und Informanten auch allein anhand der Verkehrsdaten deutlich hätte nachgezeichnet werden können – selbst bei einer Speicherfrist von zehn Wochen bzw. einer vierwöchigen bei Funkzellen. Wie sich Ausnahmeregelungen für Berufsgeheimnisträger technisch umsetzen lassen sollen, bleibt fraglich. Ein Verwertungsverbot bestimmter Kommunikationsdaten mag gut klingen – auf die Achtung der Pressefreiheit durch staatliche Behörden sollten sich Journalisten, die diese Behörden kritisieren und kontrollieren, nicht verlassen. Eine Vorratsdatenspeicherung ist eine moderne Form, journalistischen Quellenschutz zu untergraben, weshalb sich Journalisten gegen solche Vorhaben wehren sollten und das auch tun.[216]

Die Frage ist, wie lange ein solches Wehren gegen die Ausweitung staatlicher Kontrollinstrumente im Lichte der technologischen Entwicklungen noch Erfolg haben kann. Der Kulturwissenschaftler und Netzaktivist Michael Seemann hält es nur noch für eine Zeitfrage, wann die Vorratsdatenspeicherung eingeführt wird. Sie ist gewissermaßen das sichtbarste Zeichen für einen Wechsel von einer Disziplinar- in eine Kontrollgesellschaft, den Gilles Deleuze in Anlehnung an Foucault schon 1990 ausgemacht hat[217] und in der Digitalisierung ihren Höhepunkt erreicht.

> „Die Disziplinargesellschaften sind durch Foucault dadurch geprägt, dass die Individuen die gesellschaftliche Kontrolle durch ständige Disziplinierung, Überwachung und Strafe internalisieren. (...) Die Kontrollgesellschaften hingegen zeichnen sich nach Deleuze dadurch aus, dass das Regime seine Macht in entscheidungsmächtige Maschinen ausgelagert hat. In einer durchcomputerisierten Welt werden Maschinen unhintergehbar (...). Die Kontrollgesellschaft, in der jede Regung, jede Handlung, jede Entscheidung und jede Aussage von einem totalen Kontrollorgan registriert, verarbeitet und reguliert wird, ist nur die folgerichtige Vorhersage, die sich aus dem Kontrollüberschuss ergibt.“ (Seemann (2014): 75f.)

Folgt man dieser Sichtweise, bedeutet moderner Informantenschutz mehr als das Verlassen auf staatlich zugesicherte Schutzrechte. Wer Informantenschutz ernst nimmt, der versucht datensparend zu recherchieren, der bietet seinen Quellen die Möglichkeit, verschlüsselt zu kommunizieren und der bietet an, sich persönlich zu treffen, auch wenn es unbequem sein mag. Dies alles mag mühsamer sein als bisheriges Handeln. Aber es sollte zum journalistischen Selbstverständnis werden, um auf die veränderten Rahmenbedingung durch die Digitalisierung verantwortungsvoll zu reagieren.

---

[215] vgl. den Referentenentwurf „Entwurf eines Gesetzes zur Einführung einer Speicherpflicht und einer Höchstspeicherfrist für Verkehrsdaten“, abrufbar unter URL: https://netzpolitik.org/wp-upload/2015-05-15_BMJV-Referentenentwurf-Vorratsdatenspeicherung.pdf, zuletzt aufgerufen am 04.08.2015.

[216] vgl. Stellungnahme von Medienverbänden und –unternehmen zur Wiedereinführung einer Vorratsdatenspeicherung, abrufbar unter URL: http://www.djv.de/fileadmin/user_upload/Infos_PDFs/Gemeinsame_PM_11_06_15.pdf, zuletzt aufgerufen am 04.08.2015.

[217] vgl. Deleuze (1990): 254-162.

Literaturverzeichnis

BEDNER, Mark: RECHTMÄßIGKEIT DER „DEEP PACKET INSPECTION“. Projektgruppe verfassungsverträgliche Technikgestaltung (provet). Kassel 2009. Abrufbar unter URL: http://kobra.bibliothek.uni-kassel.de/bitstream/urn:nbn:de:hebis:34-2009113031192/5/BednerDeepPacketInspection.pdf, zuletzt aufgerufen am 04.08.2015.

BRANAHL, Udo: MEDIENRECHT. Eine Einführung. Wiesbaden 2013 (7. Auflage).

BUNDESMINISTERIUM DER JUSTIZ UND FÜR VERBRAUCHERSCHUTZ: ENTWURF EINES GESETZES ZUR SPEICHERPFLICHT UND EINER HÖCHSTSPEICHERFRIST FÜR VERKEHRSDATEN. Referentenentwurf (Bearbeitungsstand 15.05.2015). Abrufbar unter URL: https://netzpolitik.org/wp-upload/2015-05-15_BMJV-Referentenentwurf-Vorratsdatenspeicherung.pdf, zuletzt aufgerufen am 04.08.2015.

CHRISTL, Alexander: DATENSCHUTZ IM INTERNET. Cookies, Web-Logs, Location Based Services, E-Mail, Webbugs, Spyware. Hamburg, 2014.

DELEUZE, Gilles: POSTSKRIPTUM ÜBER DIE KONTROLLGESELLSCHAFTEN. IN: L’AUTRE JOURNAL, 1/1990, S. 254-262. Abrufbar unter: http://www.formatlabor.net/nds/Deleuze-Postskriptum.pdf, zuletzt aufgerufen am 04.08.2015.

DEUTSCHER JOURNALISTENVERBAND ET AL.: MEDIENVERBÄNDE UND – UNTERNEHMEN SAGEN NEIN ZUR VORRATSDATENSPEICHERUNG. Pressemitteilung. Berlin, 11.06.2015. URL: http://www.djv.de/fileadmin/user_upload/Infos_PDFs/Gemeinsame_PM_11_06_15.pdf, zuletzt aufgerufen am 04.08.2015.

ESCHELBACH, Ralf. In: SATZGER, Helmut/SCHLUCKEBIER, Wilhelm/WIDMAIER, Gunter: STRAFPROZESSORDNUNG. Kommentar. Köln 2014.

FRICKE, Michael/GERECKE, Martin: INFORMANTENSCHUTZ UND INFORMANTENHAFTUNG. In: AfP – Zeitschrift für Medien- und Kommunikationsrecht. Köln, 4/2014, S. 293-299.

GREENWALD, Glenn: DIE GLOBALE ÜBERWACHUNG. Der Fall Snowden, die amerikanischen Geheimdienste und die Folgen. München 2014.

HINZPETER, Britta: „COOKIE-RICHTLINIE“ IN EUROPA. Opt-in oder Opt-out – nur Deutschland legt sich nicht fest. In: Computerwoche, 07.01.2015, URL:

http://www.computerwoche.de/a/cookie-richtlinie-in-europa,2518064,2, zuletzt aufgerufen am 21.03.2015, 16:57.

HUBER, Daniel: VERWENDUNG VON COOKIES NUR NOCH BEI AUSDRÜCKLICHER EINWILLIGUNG DER NUTZER? In: IT-Recht Kanzlei München, URL: http://www.it-recht-kanzlei.de/cookies-einwilligung-datenschutz.html#abschnitt_9, zuletzt aufgerufen am 21.03.2015, 17:02.

KLAMMER, Bernd: EMPIRISCHE SOZIALFORSCHUNG. Eine Einführung für Kommunikationswissenschaftler und Journalisten. Konstanz 2005.

KOHLEN, Ralf: ROUTENIDENTIFIKATION IN VERKEHRSNETZEN AUF DER GRUNDLAGE UNSCHARFER ORTUNGSINFORMATIONEN (GSM). Dissertation. Aachen 2004.

KÖNIG, René (Hrsg.): HANDBUCH DER EMPIRISCHEN SOZIALFORSCHUNG. Stuttgart 1962.

KREMP, Matthias: DAS INTERNET JAHR 2010 IN ZAHLEN: 107.000.000.000.000 E-Mails, fast alle Spam: In: Spiegel Online, 18.01.2011, URL: http://www.spiegel.de/netzwelt/web/das-internet-jahr-2010-in-zahlen-107-000-000-000-000-e-mails-fast-alle-spam-a-740121.html, zuletzt aufgerufen am 21.03.2015, 18:31.

KROMREY, Helmut: EMPIRISCHE SOZIALFORSCHUNG. Modelle und Methoden der Datenerhebung und Datenauswertung. Opladen 1998.

KUCKARTZ, Udo: QUALITATIVE INHALTSANALYSE. METHODEN, PRAXIS, COMPUTERUNTERSTÜTZUNG. Weinheim und Basel 2012.

MACHILL, Marcel/ZENKER, Martin/BEILER, Markus: JOURNALISTISCHE RECHERCHE IM INTERNET. Bestandsaufnahme journalistischer Arbeitsweisen in Zeitungen, Hörfunk, Fernsehen und Online. Berlin 2008.

MEISTER, Andre: DEEP PACKET INSPECTION. DER UNTERSCHIED ZWISCHEN INTERNET IN DIKTATUREN UND DEUTSCHLAND IST NUR EINE KONFIGURATIONSDATEI. In: netzpolitik.org, 08.11.2012, URL: https://netzpolitik.org/2012/deep-packet-inspection-der-unterschied-zwischen-internet-in-diktaturen-und-deutschland-ist-nur-eine-konfigurationsdatei/, zuletzt aufgerufen am 04.08.2015.

NEUHÖFER, Daniel: SOZIALE NETZWERKE: PRIVATE NACHRICHTENINHALTE IM STRAFVERFAHREN. Der strafprozessuale Zugriff auf Inhalte privater Nachrichten bei Facebook & Co. In: Juristische Rundschau, Berlin, 1/2015, S. 21-31.

Literaturverzeichnis

OLTMANNS, Torsten/BRUNOWSKY, Rolf-Dieter: MANAGER IN DER MEDIEN-FALLE. Re: think CEO. Mainz 2009.

PAWLOWSKY-FLODELL, Charlotta: GRUNDLAGEN DER EMPIRISCHEN KOMMUNIKATIONSFORSCHUNG. In: JARREN, Otfried (Hrsg.): MEDIEN UND JOURNALISMUS. Eine Einführung. Opladen 1995, S. 141-170.

PÖPPELMANN, Benno H./JEHMLICH, Karina: ZUM SCHUTZ DER BERUFLICHEN KOMMUNIKATION VON JOURNALISTEN. In: AfP – Zeitschrift für Medien- und Kommunikationsrecht. Köln, 3/2003, S. 203-232.

REPORTER OHNE GRENZEN. Für Informationsfreiheit. RANGLISTE DER PRESSEFREIHEIT 2015. URL: https://www.reporter-ohne-grenzen.de/fileadmin/Redaktion/Presse/Downloads/Ranglisten/Rangliste_2015/Rangliste_der_Pressefreiheit_2015.pdf, zuletzt aufgerufen am 04.08.2015.

RICKER, Reinhardt/WEBERLING, Johannes: HANDBUCH DES PRESSERECHTS. München, 2012 (7. Auflage).

SCHLÖGL, Christian: LOGFILE- UND LINKANALYSEN. Nicht-reaktive Methoden der Online-Forschung. In: UMLAUF, Konrad/FÜHLES-UBACH, Simone/SEADLE, Michael: HANDBUCH METHODEN DER BILIOTHEKS- UND INFORMATIONSWISSENSCHAFT. Bibliotheks-, Benutzerforschung, Informationsanalyse. München, 2013, S. 184-202.

SCHMITT, Bertram. In: MEYER-GOßNER, Lutz/SCHMITT, Bertram: STRAFPROZESSORDNUNG. Kommentar. München 2014.

SCHNELL, Rainer/HILL, Paul B./ESSER, Elke: Methoden der empirischen Sozialforschung. München 2013 (10. Auflage).

SEEMANN, Michael: DAS NEUE SPIEL. Strategien für die Welt nach dem digitalen Kontrollverlust. Freiburg 2014.

SIEGERT, Paul Ferdinand: DIE GESCHICHTE DER E-MAIL. Erfolg und Krise eines Massenmediums. Bielefeld, 2008.

SOEHRING, Jörg/HOENE, Verena: PRESSERECHT. Recherche, Darstellung und Haftung im Recht der Presse, des Rundfunk und der neuen Medien. Köln, 2013 (5. Auflage).

SPITZ, Malte/BIERMANN, Brigitte: WAS MACHT IHR MIT MEINEN DATEN? Hamburg, 2014.

STÖRING, Marc: LIPPENBEKENNTNISSE AUS KARLSRUHE. Ambivalente Entscheidung zum E-Mail-Zugriff bei der Strafverfolgung. In: c't magazin, Heft 17/2009, URL: http://www.heise.de/ct/artikel/Lippenbekenntnisse-aus-Karlsruhe-292186.html, zuletzt aufgerufen am 21.03.2015, 20:35 Uhr.

Literaturverzeichnis

WALTHER, Ralf: GEOTARGETIN. Lokal gezielt besser treffen. In: SCHWARZ, Torsten: LEITFADEN ONLINE-MARKETING. Das Kompakte Wissen der Branche; Online mehr Kunden gewinnen. Waghäusel 2013, S. 338-348.

WARNKE, Martin: THEORIEN DES INTERNETS ZUR EINFÜHRUNG. Hamburg, 2011

D Code

Dieser Teil des Anhangs enthält Hinweise zu verwendetem Code, um die erlangten Befunde zu erzeugen.

D.1 Aufbereitung Daten und Erzeugung Selektoren

Der Code zur Aufbereitung der Trainingsdaten und zur Selektorgenerierung befindet sich in den abgegebenen Dateien zur Masterarbeit im `code/`-Ordner.

D.1.1 “Own”-Datenformate nach file(1)-Analyse: Top 50 (Originaldateien)

```
cat files+types_1_raw.txt | cut -d ":" -f2 | sed -e "s|^ ||g" | sed -e "s|,.*||g" | sort | uniq -c | sort -gr | head -50
 124505 ASCII text
  91210 data
   9527 HTML document
   9207 UTF-8 Unicode text
   8988 GIF image data
   6051 JPEG image data
   3000 PNG image data
    471 XML document text
    406 ISO-8859 text
    378 very short file (no magic)
    269 exported SGML document
    230 SPARC demand paged executable not stripped
    225 SPARC demand paged executable
    216 SVG Scalable Vector Graphics image
    167 RIFF (little-endian) data
    166 Web Open Font Format
    155 Non-ISO extended-ASCII text
    104 MS Windows icon resource - 1 icon
     97 Debian binary package (format 2.0)
     69 bzip2 compressed data
     67 C source
     51 assembler source
     45 MGR bitmap
     43 MS Windows icon resource - 2 icons
     43 Emacs v18 byte-compiled Lisp data
     42 UTF-8 Unicode (with BOM) text
```

```
     35 XZ compressed data
     33 PDF document
     31 Macromedia Flash data (compressed)
     23 MIME entity
     18 XML 1.0 document text
     18 TrueType font data
     15 Zip archive data
     15 MS Windows icon resource - 4 icons
     13 troff or preprocessor input
     13 PGP public key block Public-Key (old)
     12 MS Windows icon resource - 3 icons
     10 gzip compressed data
     10 C++ source
      8 DOS executable (COM)
      7 PC bitmap
      6 Minix filesystem
      6 8086 relocatable (Microsoft)
      5 PGP signature Signature (old)
      4 WebM
      4 Pascal source
      4 MS Windows icon resource - 6 icons
      4 Embedded OpenType (EOT)
      3 PGP\011Secret Sub-key -
      3 OpenStreetMap XML data
```

D.1.2 “Own”-Datenformate nach file(1)-Analyse: Top 50 (Keine Duplikate)

```
cat files+types_2_raw_uniq.txt | cut -d ":" -f2 | sed -e "s|^ ||g" | sed -e "s|,.*||g" | sort | uniq -c | sort -gr | head -50
  97410 ASCII text
  89873 data
   6918 HTML document
   6161 UTF-8 Unicode text
   5540 JPEG image data
   2341 PNG image data
   1092 GIF image data
    394 ISO-8859 text
    284 XML document text
    245 exported SGML document
    230 SPARC demand paged executable not stripped
    225 SPARC demand paged executable
    158 SVG Scalable Vector Graphics image
    154 Non-ISO extended-ASCII text
    141 RIFF (little-endian) data
    111 Web Open Font Format
     97 Debian binary package (format 2.0)
     86 MS Windows icon resource - 1 icon
     52 bzip2 compressed data
     44 assembler source
     40 C source
     35 Emacs v18 byte-compiled Lisp data
     33 PDF document
```

```
     31 XZ compressed data
     30 MS Windows icon resource - 2 icons
     29 UTF-8 Unicode (with BOM) text
     28 Macromedia Flash data (compressed)
     23 MIME entity
     21 MGR bitmap
     17 very short file (no magic)
     16 TrueType font data
     13 PGP public key block Public-Key (old)
     12 XML 1.0 document text
     10 Zip archive data
     10 MS Windows icon resource - 3 icons
     10 gzip compressed data
      8 troff or preprocessor input
      8 DOS executable (COM)
      7 MS Windows icon resource - 4 icons
      7 C++ source
      6 PC bitmap
      6 Minix filesystem
      6 8086 relocatable (Microsoft)
      5 PGP signature Signature (old)
      4 Pascal source
      4 Embedded OpenType (EOT)
      3 PGP\011Secret Sub-key -
      3 OpenStreetMap XML data
      3 MS Windows cursor resource - 1 icon
      3 COM executable for MS-DOS
```

E Daten

Hier bestehen Hinweise zu Trainings- und Evaluationsdaten, die in dieser Arbeit genutzt wurden.

E.1 Trainingsdaten

Trainingsdaten finden sich in verschiedenen Verarbeitungsstufen unter `data/train/`.

E.2 Evalautionsdaten

Die Evaluationsdaten, die in der Arbeit eingsetzt wurden, sind unter `eval/` zu finden: diese können allerdings aus Urheberrechtsgründen (Filme, Bücher und weitere Werke sind darunter) nicht vollständig veröffentlicht werden, sind auf Anfrage aber einsehbar.

F Hashsummen

Es werden die SHA512-Hashsummen gemäss offizieller Abgabe vom 30.11.2015 – mittels *sha512sum(1)*-Kommando ermittelt – zur Verfügung gestellt.

F.1 Code

Für `code.tar`:

28293c59c988b4a33f028be427a8251c91caea4c2e4aa1893db5ce9e6542ffd93f76a0e693ca4d996540aa3b2b11c899b716fa72a8406b3b5f639ceb2d404a58

F.2 Daten

Für `data_public.tar`:
a651d2d225ae9da3e95f8a1a1422f3b3b15810e15dbd03bf37567c965af2fa6a26dec9fd125e3c50216ae5bf89bf391800a5ece8a696faef2bd4afccb563c06c

Für `eval_public.tar`:
b2de2e4b41294664a527c9aa1a470f8514b19ffbba14ad1f0640282aea9a56541e15f9343ae8c6dd5983b14f4414e4a6e598a77578b0a821a5ee9d5302fd5d6d